Trinity College Dublin
Coláiste na Tríonóide, Baile Átha Cliath
The University of Dublin

Vertical Governance of Online Speech: Evidence from Google's Moderation Mandate

Mike McRae

TEP Working Paper No. 1425

October 2025

Abstract:

This paper provides the first causal evidence that upstream infrastructure providers can reshape social media discourse by enforcing content moderation through access-based leverage. I exploit a 2022 update to Google's Play Store policy requiring stricter removal of violent threats and misinformation, along with variation in platform exposure across three similar `Alt-Tech' social media platforms, within a triple-differences design...

Vertical Governance of Online Speech: Evidence from Google's Moderation Mandate

Mike McRae*

October 15, 2025

Abstract

This paper provides the first causal evidence that upstream infrastructure providers can reshape social media discourse by enforcing content moderation through access-based leverage. I exploit a 2022 update to Google's Play Store policy requiring stricter removal of violent threats and misinformation, along with variation in platform exposure across three similar 'Alt-Tech' social media platforms, within a triple-differences design. Using a novel panel of over 28 million posts from 62,000 users, I find that threatening content declined sharply and persistently on the exposed platforms, particularly among high-risk users. These effects are not explained by user self-censorship, public awareness, contemporaneous events, or selective data loss. I also document significant reductions in lawful but politically sensitive narratives, including election denial and January 6 insurrection commentary. The findings show how infrastructure-level enforcement can durably alter the boundaries of permissible speech across platforms, contributing to literatures on platform governance, vertical restraints in digital markets, and the institutional foundations of online discourse.

Keywords: platform governance; content moderation; digital infrastructure.
JEL Codes: D83, L86, L82.

*Trinity College Dublin. I thank Carlo Schwarz, Nicola Mastrorocco, Enrico Cavalotti, Nicola Fontana, Elliot Motte and Tommaso Colussi for helpful feedback and suggestions. I also thank all participants at the 2nd Text-as-Data in Economics Workshop at Lancaster University for informal discussions. All errors are my own.

1 Introduction

In digital markets, a small number of private infrastructure providers control the primary channels through which users access applications, information, and communication tools. While governments continue to regulate online speech to varying degrees, the actions of private infrastructure firms such as mobile app stores and cloud service networks have the potential to reshape the boundaries of discourse in ways that are neither fully transparent nor democratically accountable.

While prior work has shown that stronger moderation policies on social media platforms can reduce harmful content (Allcott and Gentzkow, 2017; Klonick, 2018; Roberts, 2019), little is known about how such moderation is shaped by upstream infrastructure providers acting through market access conditions. This paper studies one such enforcement mechanism: a dominant mobile app distributor's imposition of stricter moderation requirements on platforms. In July 2022, Google announced a new policy requiring that all apps on the Play Store implement robust systems to remove threats of violence and misinformation related to sensitive events, such as public health emergencies or allegations of electoral fraud. Apps perceived as endorsing or profiting from denial of such events could be suspended. Importantly, the policy applied both to existing apps and to those seeking admission but not to apps with no ties to the Google marketplace, generating a quasi-experimental setting in which platforms faced differential exposure to the new enforcement regime.

This dynamic reflects a form of vertical control in which upstream intermediaries impose behavioural constraints on downstream platforms through access-based leverage rather than ownership or pricing. In contrast to classical models of vertical integration or exclusive dealing (Rey and Tirole, 2007; Lafontaine and Slade, 2007; Rey and Vergé, 2004; Hart and Tirole, 1990; Klein and Murphy, 1988), the enforcement studied here operates through infrastructure bottlenecks in digital distribution, allowing dominant firms to condition downstream conduct by controlling access to users. Much like franchisors enforcing uniform service standards or retailers requiring quality certification for shelf space, these non-price vertical restraints shape platform behaviour without formal ownership or pricing mechanisms. This logic is especially salient in two-sided markets (Rochet and Tirole, 2003; Armstrong, 2006), where platforms must balance user engagement and advertiser monetisation while remaining compliant with upstream distribution constraints. This paper provides the first causal evidence that vertical governance, exercised through mobile distribution control, can reshape platform moderation.

Identifying the causal effect of moderation policies on user content is challenging, as platform decisions to moderate are typically endogenous to factors such as ideology, user base composition, and growth strategy. Assessing the role of upstream enforcement adds further difficulty due to the opacity of policy implementation and the prevalence of 'deplatforming'[1], which often creates discon-

[1]The practice of removing apps from the digital infrastructure entirely

tinuities in data and induces user migration (Agarwal et al., 2022). I leverage a sharp and largely unpublicised shift in Google's developer policy that, crucially, did not result in any 'deplatforming' and, to the best of my knowledge, received no media coverage. This setting enables continuous observation of content across platforms without confounding spillovers from public awareness or shocks to the overall availability of the platforms. My identification strategy exploits variation along three dimensions: platform exposure to Google's policy, user-level pre-treatment threat intensity, and time; within a triple-differences framework. Specifically, I compare changes in threatening content between high- and low-risk users before and after the policy across treated platforms (Gettr and Truth Social) and a structurally untreated counterfactual (Gab), which remained excluded from the Play Store throughout.

The analysis draws on a novel panel dataset of over 28 million user posts from 62,000 users across the three platforms, spanning February 2021 to December 2024. I measure the 'threateningness' of posts using scores from the Perspective API (Wulczyn et al., 2017; Dixon et al., 2018) and track the prevalence of content categories explicitly targeted by the policy, such as election denial, anti-vaccine misinformation, and discourse related to the January 6 Capitol insurrection.

To validate this design, I begin by showing that Gab serves as a credible counterfactual, assessing pre-treatment linguistic similarity and parallel trends in outcome evolution both within and across platforms. I then systematically address several additional identification threats.

First, I rule out mechanical explanations. Although the policy applied only to Android apps, users post across platforms, and my crawler captures only web-visible content. The sharp drop in threatening posts at the treatment date is inconsistent with Android-only or retroactive enforcement, implying a conservative estimate of the true effect. Second, I address concerns about sample selection and attrition. High-threat users are more likely to remain active, especially on Gab, which could attenuate moderation effects. However, treatment effects remain stable across both balanced and unbalanced samples, suggesting that panel composition does not drive the results. Third, I examine selection into platforms. If users with higher threat levels sorted disproportionately into treated or untreated platforms before the policy, platform differences could reflect composition rather than treatment. This risk is mitigated by the DDD design, which includes user fixed effects and relies on within-user changes over time. I further confirm overlapping pre-treatment threat distributions across platforms and replicate the main results using matched subsamples. Fourth, I rule out alternative explanations tied to contemporaneous U.S. events (e.g., the Mar-a-Lago raid, midterm primaries, or Apple policy updates) through dynamic placebo tests and a cross-country falsification analysis using a large sample of Brazilian users. Fifth, I test for user self-censorship and find no evidence of anticipatory behavior, no detectable changes in platform moderation policy language, and treatment effects that increase monotonically with user-level risk, especially in the immediate aftermath of the policy. These patterns are inconsistent with purely user-driven adjust-

ment. Finally, I test for spillovers. Among users active on both Gettr and Gab, I find no evidence that exposure to treated platforms altered behavior on Gab or disrupted aggregate trends.

Turning now to the main estimation, results show that a one standard deviation increase in a user's pre-policy level of threatening content reduced such content by 40–60 percent of the pre-treatment mean on Gettr and Truth Social relative to Gab, with effects emerging immediately after the policy change and persisting over the entire study period, the end of 2024. Results are monotonically increasing in user 'threateningness' and are robust to user fixed effects, platform-specific time trends, and a wide range of alternative outcome and user exposure definitions. Taken together, the evidence supports a causal interpretation: upstream enforcement by Google induced measurable and durable changes in downstream moderation on exposed platforms.

Beyond the moderation of threatening content, I also document a distinct and independent shift in the prevalence of politically sensitive narratives, including election fraud claims, anti-vaccine discourse, and January 6 commentary. Users who were previously active in discussing these topics significantly reduced such engagement following the policy change. Given the timing and consistency of these declines across affected platforms, the evidence, again, points to platform-enforced moderation as the most plausible mechanism. This broader contraction of salient but largely lawful discourse highlights how infrastructure-based enforcement can compel platforms to reshape the contours of acceptable speech, not only by suppressing unlawful or violent content but by dampening controversial narratives prevalent in the ecosystem. These effects highlight how vertical governance by upstream infrastructure providers can significantly alter the landscape of discourse across the entire digital ecosystem, superseding fragmented or inconsistent platform-level rules.

A central concern in platform governance is whether content moderation suppresses harmful behaviour or simply pushes it elsewhere. Prior studies have shown that public, high-profile moderation efforts, such as coordinated bans or app store removals, often trigger backlash, user migration, and content reconstitution on less regulated platforms (Rizzi, 2024; Agarwal et al., 2022; Horta Ribeiro et al., 2023). To assess whether Google's more silent, infrastructure-level policy shift produced similar spillover dynamics, I examine three channels of behavioural adjustment. First, I test whether threatening users disengaged or reduced visibility on treated platforms. Second, I evaluate whether users reacted to perceived enforcement by discussing moderation before exiting. Third, I assess whether users displaced threatening content to Gab after leaving Gettr. Across all three domains, the evidence points to internal adjustment within treated platforms, not reactive redistribution or cross-platform reallocation. This suggests that infrastructure enforcement can reduce harmful content without triggering the backlash dynamics observed in more visible moderation episodes.

The findings point to several important consequences for platform governance. First, control over distribution can translate into control over what information circulates online. When a single

infrastructure provider conditions access to the market, it effectively sets rules that platforms must follow, even in the absence of formal regulation. Second, enforcement actions by one distributor often lead platforms to implement policy changes globally, affecting all users regardless of how they access the platform. Third, this asymmetry creates incentives for free-riding by other distributors who benefit from enforcement without bearing its costs. It also complicates decision-making for platforms that must navigate conflicting expectations across different gatekeepers. Finally, the combination of economic intermediation and content-based enforcement blurs the traditional line between market infrastructure and regulatory authority. This calls for a policy framework that takes seriously the role of infrastructure providers in shaping online speech.

These dynamics raise broader concerns about legitimacy, accountability, and welfare in digital speech markets. While private enforcement can reduce harmful content, it also shifts the power to define acceptable expression from public institutions to private firms. In this case, Google was able to suppress a wide range of political narratives that were neither illegal nor factually false. These were precisely the types of speech that alternative platforms were designed to support. Yet the threat of losing market access was enough to remove them from the ecosystem. If a small number of infrastructure providers can reshape platform behaviour through commercial pressure alone, the marketplace of ideas may reflect the risk preferences of a few firms rather than democratic or legal standards. The result may be excessive caution, leading platforms to over-comply and suppress controversial but lawful discourse. Understanding how these governance dynamics operate through market structure is essential for assessing the broader effects of infrastructure control on digital expression.

This paper contributes to three strands of literature. First, it adds to the growing body of work on content moderation and governance in digital media platforms. Existing research has examined how platforms regulate speech internally, balancing user engagement with reputational, legal, and normative pressures (Liu et al., 2021; Madio and Quinn, 2024; Kominers and Shapiro, 2024; Beknazar-Yuzbashev et al., 2024). Scholars have shown that platform architecture and recommendation systems shape users' exposure to misinformation and political polarisation (Allcott et al., 2020; Levy, 2021), and influence levels of hate speech, incivility, and partisan sorting (Mosquera et al., 2020; Müller and Schwarz, 2021; Cao et al., 2023). Recent work by Kalra (2025) finds that depersonalised content feeds reduce toxicity but at the cost of user engagement. Within this literature, one prominent line of empirical research focuses on platform self-governance. For example, Müller and Schwarz (2023) document reductions in toxic language after Twitter removed Donald Trump's account, while Rizzi (2024) shows that Twitter's hate speech policy reduced hate on-platform but induced cross-platform spillovers to Parler. A related stream examines the role of public regulation, such as Germany's NetzDG law, which has been shown to reduce hate speech and associated offline violence (Duran et al., 2022; Andres and Slivko, 2021).

A closely related but still nascent literature examines infrastructure-level content governance, where upstream intermediaries, such as app stores or cloud hosts, shape downstream platform behaviour by conditioning market access. Existing work focuses primarily on deplatforming events, showing that app store removals can trigger user migration to fringe alternatives with higher levels of misinformation and toxicity (Agarwal et al., 2022; Horta Ribeiro et al., 2023). However, these studies leave open whether infrastructure pressure alters behaviour within platforms, rather than simply shifting users across them. This paper addresses that gap by providing the first causal evidence that vertical enforcement by infrastructure providers can induce moderation changes within affected platforms.

Moreover, unlike prior deplatforming episodes that triggered measurable spillovers to alternative platforms (Rizzi, 2024; Agarwal et al., 2022; Horta Ribeiro et al., 2023), I find no evidence of large-scale redistribution of harmful content to less regulated environments. Cross-platform users did not significantly increase threatening discourse on Gab following exit from moderated platforms, and moderation-related discourse rose steadily across all platforms, suggesting broad and diffuse engagement with content rules rather than concentrated backlash. This distinction highlights the importance of enforcement visibility: by operating silently and within the infrastructure, Google's policy appears to have reduced harmful content without provoking mass migration or content displacement, a dynamic that challenges conventional narratives of reactive user spillover. In doing so, the paper reframes infrastructure intermediaries not just as gatekeepers of market access, but as de facto regulators capable of shaping online discourse at scale. This advances the literature by uncovering a previously under-explored channel of governance, non-price vertical control, with implications for the political economy of digital speech, platform competition, and regulatory design.

This paper also contributes to the industrial organisation literature on vertical restraints in digital markets. Classic models emphasise how upstream firms can influence downstream behaviour through non-price mechanisms such as shelf-space access, franchising, or certification requirements (Klein and Murphy, 1988; Rey and Tirole, 2001; Lafontaine and Slade, 2007). I extend this logic to the governance of political speech in digital platforms, showing how mobile app stores act as infrastructure bottlenecks that condition downstream behaviour without direct ownership or pricing. In doing so, the paper builds on recent work by Madio et al. (2025), who document how financial intermediaries enforce content standards in adult entertainment. My findings generalise this mechanism to a broader regulatory context, demonstrating how vertical control operates in mainstream political discourse and scales across platform architectures.

The third and final strand relates to the institutional foundations of political communication and digital governance. Foundational work in political economy shows how institutional actors influence the flow of information and political accountability, often via control over media ownership or content distribution (Besley and Prat, 2006; Prat, 2018). I extend this logic to infrastructure

providers who, while not themselves content producers or governments, exercise substantial de facto authority over discourse by determining which platforms can access distribution channels and under what conditions. The enforcement regime I study thus resembles the kind of informal institutional power described by Acemoglu and Robinson (2008). Digital infrastructure, like legacy media systems, constitutes a locus of political power, capable of shaping expressive incentives and altering the effective rules that govern online speech.

This paper is structured as follows. Section 2 provides background and institutional details of mobile app distribution, content-moderation policies, and the growth of alternative platforms. Section 3 describes the data. Sections 4 discusses identification and explores the causal effect of the Google announcement on online behaviour. Section 5 discusses policy implications and Section 6 concludes.

2 Background

2.1 Private Governance, Infrastructure, and the Rise of Alt-Tech Platforms

In recent years, social media platforms have come under increasing pressure to moderate harmful or extreme content, including hate speech, misinformation, and incitements to violence. These pressures originate from governments, advertisers, civil society groups, and users, and reflect mounting concern about the downstream consequences of digital speech (Muller and Schwarz, 2019; Duran et al., 2022). In some settings, governments have implemented legal frameworks requiring platforms to remove specific types of content, Germany's 2017 Network Enforcement Act (*NetzDG*) being the most prominent example (Duran et al., 2022). But outside a few jurisdictions, content moderation remains primarily a matter of private governance, shaped by reputational concerns, platform policies, and advertiser preferences.

One consequence of this evolving regulatory environment has been the emergence of a parallel ecosystem of alternative social media platforms, often referred to as "Alt-Tech."[2] These platforms, including Truth Social, Gab, Gettr, and Parler, among others, have positioned themselves as ideological counterweights to mainstream platforms, promising to uphold "free speech" and resist what they characterise as politically biased or excessive moderation. Their appeal lies in their willingness to host content disallowed elsewhere, and their user bases often include groups disillusioned with

[2]The term "alt tech" is a portmanteau of "alt right" and "Big Tech," a phrase adopted by some prominent conservatives and their supporters beginning in 2015, following bans from mainstream platforms. Roose, Kevin. "The Alt-Right Created a Parallel Internet. It's an Unholy Mess." *The New York Times*, December 11, 2017 (`https://www.nytimes.com/2017/12/11/technology/alt-right-internet.html`).

the enforcement policies of larger platforms. Empirical research has documented the prevalence of polarising, conspiratorial, or hateful content on these sites (Zannettou et al., 2018; Aliapoulios et al., 2021), emphasising the challenges of moderating harmful speech in a fragmented online environment.

Although these platforms market themselves as unregulated spaces, their viability depends on access to key infrastructure. Many operate under business models reliant on advertising or subscriptions, which require sustained engagement and a sizeable user base, conditions that, in practice, often depend on discoverability through app stores and accessibility across devices. To reach users at scale, Alt-Tech platforms rely on third-party services for web hosting, payment processing, and mobile app distribution, dependencies that expose them to external forms of governance.

A growing number of infrastructure providers have adopted terms of service prohibiting the facilitation of hate speech or violent content, and have exercised this authority to deny access to noncompliant platforms. Parler, for example, was removed from both the Apple and Google app stores in January 2021 and later reinstated on Apple following adherence to moderation policies.[3] Gab was delisted from both major app stores in 2017. While these are headline Alt-Tech examples, smaller-scale enforcement is commonplace for all apps. Developer forums such as the iOS Developer and Android Developer subreddits contain numerous reports of apps being rejected or removed for failing to adhere to moderation-related policies, particularly those governing user-generated content.[4]

In addition to app store removals, infrastructure enforcement has taken several forms. For example, Amazon Web Services suspended Parler's hosting in January 2021, rendering the site inaccessible until it migrated to alternative providers. Cloudflare has withdrawn Distributed Denial of Service (DDoS) protection from multiple fringe platforms, including The Daily Stormer (2017), 8chan (2019), and Kiwi Farms (2022), each time citing violations of anti-abuse or hate speech policies. Payment processors such as PayPal and Stripe also cut off services to Gab in 2018 after it was linked to extremist violence. A full timeline of these and a non-exhaustive list of other examples is provided in Appendix Table C.1.

These cases illustrate the potential for infrastructure providers to exert meaningful control over platform access, not by removing content directly, but by conditioning the availability of critical technical services on compliance with moderation standards. Among these intermediaries, mobile app stores occupy a particularly influential position. As smartphone usage has become ubiquitous, most users now access social media platforms primarily through mobile apps rather than web browsers.[5] Two firms, Apple and Google, control the only official app stores available on iOS

[3]Clayton, James. 2021. "Parler set to return to Apple's App Store." *BBC News*, April 19, 2021

[4]`https://www.reddit.com/r/iOSProgramming/`; `https://www.reddit.com/r/androiddev/`

[5]According to Statista, over 95% of U.S. social media users access platforms primarily via mobile apps.

and Android devices, respectively, and require that platforms comply with a range of technical and content-related standards in order to remain listed. These policies prohibit the promotion of violence, hate speech, and other forms of harmful content, and, as mentioned, have been cited in multiple app removals or rejections.

2.2 Google's 2022 Enforcement Shock

On July 27, 2022, Google updated its Developer Program Policy to clarify and expand its expectations under the "Inappropriate Content" clause. It reiterated Google's stance against apps containing threatening or violent content and introduced a revised provision on "Sensitive Events" (see Figure C.2). The update prohibits apps from capitalizing on or being insensitive toward events with "significant social, cultural, or political impact," including civil unrest, public health emergencies, natural disasters, and deaths. While content related to such events may be allowed if it holds educational, documentary, scientific, or artistic (EDSA) value, apps must not, for example, deny the occurrence of "well-documented, major tragic events" or appear to profit from such events without discernible benefit to victims. Compared to the previous version, the revised policy broadened the scope of sensitive events and made enforcement language more precise. Although framed as a clarification, the update was immediately enforceable and applied to both new and existing apps without a grace period.[6] This shift has important implications for user-generated content platforms, which may now face heightened liability not only for individual posts but for broader patterns of content or monetisation. Platforms hosting controversial or conspiratorial speech must implement more robust moderation systems to assess factual accuracy, contextual sensitivity, and commercial exploitation, raising compliance costs and the likelihood of content takedowns. The July update also included stricter provisions on misleading health claims, impersonation, ad placement, subscription cancellation, and VPN usage.[7]

This update was part of a broader sequence of policy revisions across 2022 that progressively expanded the scope and specificity of Google's content moderation requirements. On April 6, Google announced revised rules for user-generated content (UGC), scheduled to take effect on October 11, which placed system-level responsibility on platforms to implement proactive and ongoing moderation (see Figure C.3). Google also issued a clarification of its hate speech policy, reaffirming its prohibition on incitement and outlining expectations for compliance with local legal standards in different jurisdictions.[8]

See: `https://www.statista.com/topics/4689/mobile-social-media-usage-in-the-united-states`

[6] Archived version retrieved from the Internet Archive: `https://web.archive.org/web/20220727163419/https://support.google.com/googleplay/android-developer/answer/12253906`

[7] Some categories had specific deadlines, e.g., 30 days or a date within three months; the provisions relevant to this study were immediate.

[8] See `https://web.archive.org/web/20220406190711/https://support.google.com/googleplay/`

By late July, enforcement activity had intensified, and the immediate applicability of the sensitive events clause introduced a new source of compliance exposure for platforms operating in sensitive political or social domains. A rejection of Truth Social from the Play Store on August 19, citing inadequate moderation systems and threatening content, which is discussed in the proceeding section, provided clear evidence that monitoring and enforcement was already being carried out in practice. Table A.1 summarises the key policy milestones over this period.

Unlike other provisions introduced earlier in the year, which included future compliance deadlines, the sensitive events clause took effect without delay. This makes July 27 a natural threshold for treatment exposure. From that point forward, platforms distributing on Android were subject to heightened enforcement risk and regulatory scrutiny. The empirical analysis that follows adopts this timing convention, comparing treated and untreated platforms to assess how this shift in enforcement salience influenced strategic moderation responses.

2.3 Gettr - Existing Google Play App

Gettr is a social media microblogging platform operating in the Alt-Tech space, presenting itself as a free speech alternative to mainstream platforms. Launched on July 4, 2021, by former Trump administration spokesperson Jason Miller, it quickly attracted a user base composed largely of conservatives and others who viewed traditional platforms as overly restrictive or politically biased in their moderation practices. While Gettr promised fewer restrictions on speech, it also emphasised a baseline commitment to user safety. As of early 2022, the app had been downloaded 6.5 million times globally across the Apple App Store and Google Play.[9] When Google's 2022 policy update introduced stricter content moderation requirements for apps distributing user-generated content, Gettr, as an existing Play Store application operating in a politically sensitive niche, faced heightened compliance expectations despite the absence of any publicly disclosed sanctions.

The potential removal of Gettr from the Google Play Store would entail substantial costs across multiple dimensions. First, from a user acquisition standpoint, the Play Store serves as a primary gateway for Android users to discover and install applications.[10] De-platforming would necessitate reliance on alternative distribution methods, such as direct APK downloads or third-party app stores, which historically have shown significantly lower adoption rates due to user friction and

android-developer/answer/11899428

[9]David Klepper and Barbara Ortutay, "A Year After Trump Purge, 'Alt-Tech' Platform Offers Far-Right Refuge," *Associated Press*, February 5, 2022. https://apnews.com/article/donald-trump-technology-business-social-media-joe-rogan-83bdedc0163473e393918f95f889d594

[10]The Play Store provides a central platform for users to browse, search, and find apps. It's the official and pre-installed app store for Android devices, and includes features which ensures apps are kept up-to-date with the latest versions and security patches, making it the most reliable and secure way to get apps.

security concerns (Goodwin and Woolley, 2022). Second, user engagement and retention could suffer; without access to the Play Store, users would not receive automatic updates, leading to a fragmented user experience and potential security vulnerabilities. Third, reputational risks are considerable. The removal of an app from major platforms often garners media attention and can be perceived as a signal of non-compliance or association with harmful content, potentially deterring both users and advertisers. The case of Parler illustrates these challenges: following its removal from the Google Play Store in January 2021 due to inadequate moderation policies, Parler experienced a significant decline in daily active users and faced difficulties in maintaining its infrastructure and user base.[11] [12]

2.4 Truth Social - Prospective Google Play App

Truth Social is a similar Alt-Tech micro-blogging platform, launched on February 21, 2022, by Trump Media & Technology Group (TMTG), a company founded by U.S. President Donald Trump. Marketed as an alternative to mainstream social media platforms, the platform positioned itself as a "haven of free speech," explicitly catering to users who believed major tech companies unfairly censored conservative viewpoints. Truth Social's launch followed Trump's permanent suspension[13] from Twitter in January 2021, which spurred efforts to create an independent platform free from the content moderation policies of Big Tech.

Although Truth Social was available to iOS users from its launch and accessible via web browser, it was initially unavailable on Google Play. As noted in Section 2, Google's new policy explicitly required all user-generated content platforms, whether existing or seeking listing, to demonstrate "effective systems for moderating" sensitive or violent content. On August 19, 2022, Google notified TMTG that Truth Social's Android app submission violated these requirements, reportedly sharing screenshots of posts that incited physical violence. In a statement reported at the time, a Google spokesperson reiterated that an effective moderation system was a condition for any app to be listed on Google Play. Truth Social responded that it was working to address these issues and was committed to compliance "without compromising our promise to be a haven for free speech."[14]

The incident came at a sensitive period for TMTG, whose planned merger with Digital World

[11] Chen, S. (2021, January 9). *Google suspends Parler from app store after deadly Capitol violence.* Axios. Retrieved from https://www.axios.com/2021/01/09/capitol-mob-parler-google-ban

[12] Hosseinmardi, H., Ghasemian, A., Clauset, A., Mobius, M., Rothschild, D., & Watts, D. J. (2023). *Deplatforming did not decrease Parler users' activity on fringe social media.* PNAS Nexus, 2(3), pgad035. https://doi.org/10.1093/pnasnexus/pgad035

[13] This suspension was lifted in November 2022, following Elon Musk's acquisition of Twitter. See https://www.bbc.com/news/world-us-canada-63692369.

[14] Nico Grant, "Google Says Trump's Truth Social Must Scrub Violent Content to Join Play Store," *The New York Times*, August 30, 2022.

Acquisition Corp. (a SPAC that had raised nearly $300 million) was under federal investigation. Delays in app store approval threatened the platform's growth and investor confidence.[15] On October 12, 2022, Google reversed its initial decision and approved Truth Social for distribution on Google Play, contingent on its adherence to the same requirement to remove incitements to violence. This listing significantly expanded Truth Social's reach among Android users, who represent roughly 42% of the U.S. smartphone market.[16]

By contrast with Gettr, already on Google Play and never facing a public enforcement announcement, Truth Social underwent a period of outright exclusion. Other Alt-Tech platforms, such as Rumble, had already secured placement on Google Play, while Parler was in the process of regaining access,[17] creating further pressure for Truth Social to comply.

While Truth Social publicly maintained its ideological commitments, its eventual accommodation of Google's moderation requirements raises a broader question: did the threat of exclusion from a key distribution channel induce substantive changes in platform behaviour? Some reporting suggests that Truth Social may already have had internal moderation infrastructure in place.[18] Whether these actions reflected an internal policy shift, an effort to preempt enforcement, or a response to other pressures remains unclear.

2.5 Gab - Google Play Excluded

Gab serves as a critical test case for distinguishing between Google-led enforcement effects and broader ecosystem-wide changes in user behaviour or platform moderation. Founded in 2016, Gab was created as a free speech-oriented alternative to mainstream platforms, explicitly targeting users who felt marginalised by existing content moderation practices. The platform features a microblogging format similar to Truth Social and Gettr, with short posts ("gabs"), reposts, and a follower network.

Gab officially launched on Google Play in May 2017 but was removed by Google on August 17, 2017, for failing to demonstrate a sufficient level of moderation, including for content that encourages violence and advocates hate against groups of people."[19] In response, Gab filed an antitrust lawsuit

[15] See Matthew Goldstein, "Google Approves Trump's Truth Social for Its Play Store," *The New York Times*, October 13, 2022. `https://www.nytimes.com/2022/10/13/business/truth-social-google-digital-world.html`

[16] `https://gs.statcounter.com/os-market-share/mobile/united-states-of-america/`.

[17] https://www.pewresearch.org/journalism/2022/10/06/alternative-social-media-appendix-a-detailed-tables-for-audit-of-seven-alternative-social-media-sites/

[18] Anecdotal evidence from mid-2022 indicates the removal of content critical of Donald Trump. See https://www.vanityfair.com/news/2022/06/truth-social-moderation-trump

[19] See Google faces lawsuit over removing Gab from Play Store," *BBC News*, September 18, 2017. https://www.bbc.com/news/technology-41306437

against Google, which it dropped in October 2017 in favor of lobbying efforts targeting alleged monopolistic practices by Big Tech.[20]

Unlike other platforms in the Alt-Tech ecosystem, Gab has maintained an explicit ideological commitment to resisting external pressures to moderate user content. Its leadership has consistently framed content moderation requirements by major platforms and app stores as violations of free speech principles.[21] As a result, Gab has relied primarily on web-based access and third-party installations, significantly altering its infrastructure compared to peers like Gettr and Truth Social that remained dependent on app store distribution. Gab earns revenue through premium subscriptions, donations, and affiliate advertising.[22]

Despite these differences in infrastructure and access, Gab's user base and ideological positioning are highly similar to those of Gettr and Truth Social. This similarity makes Gab an important counterfactual: because it was already excluded from Google Play prior to the 2022 policy update and had no ambition of rejoining the distribution network, any subsequent changes in user behaviour or moderation practices on Gab cannot plausibly be attributed to new distribution-related pressures. Comparing Gab to other Alt-Tech platforms thus helps isolate the role of app store enforcement from broader endogenous shifts in the Alt-Tech ecosystem.

3 Data

My analysis draws on newly constructed datasets from three social media platforms: Truth Social, Gettr, and Gab. I assemble detailed post-level, comment-level, and user-level datasets from each platform to study the evolution of threatening and 'senseitive' content and platform level responses to Google's 2022 content moderation policy update.

3.1 Truth Social, Gettr, and Gab

I collect comprehensive user activity data from the three studied platforms, capturing the full period from each platform's public launch through December 2024. Specifically, the data begin in July 2021 for Gettr (its launch month) and in February 2022 for Truth Social (following its release), ensuring complete coverage from inception. I assume that content moderation is implemented platform-wide and collect data from the publicly accessible web interfaces of each platform. Of course, if

[20] See ibid.

[21] Coldewey, D. (2017, August 17). *Alt-social network Gab booted from Google Play Store for hate speech.* TechCrunch.

[22] See Gab 2020 SEC Annual Report Filing

changes in moderation were implemented only for Android-based app users (e.g., in response to Google Play policy), such changes would not appear in my dataset, an important consideration when interpreting platform-wide effects.

To ensure comparability across platforms, I apply a consistent sampling frame. I first compile extensive lists of known user accounts for each platform, then randomly sample users who were active in both the pre- and post-policy periods. This ensures that the final samples focus on users with sustained engagement, making behaviour comparable across platforms and over time. Table A.3 provides detailed summary statistics for the samples used, including user counts, post counts, timelines of data collection, and descriptive statistics for engagement measures. The estimates show broadly similar post-level characteristics across platforms in the pre-treatment period. Average threat scores and incidence of key topic flags (e.g., election fraud, anti-vaccine, QAnon) are of comparable magnitude, suggesting that the baseline distribution of harmful or conspiratorial content is relatively consistent. One notable exception is the substantially higher engagement on Truth Social which likely reflects dynamics during the platform's onboarding phase. This may be partly organic, given concentrated early activity among prominent users, but could also reflect developer-side amplification or algorithmic design aimed at boosting visible engagement, a pattern observed historically during the growth phases of some social platforms.[23]

For each platform, I begin with a pool of approximately 2 million known user accounts from which I randomly sample users. These user lists are assembled by seeding with a randomly selected user and recursively collecting all followers and followings, continuing the process until the network expansion yields no new users for an extended period. This approach produces a near-complete snapshot of the active user network at the time of collection, excluding only those accounts that were closed, suspended, or otherwise inaccessible. I then randomly select 20,000 users, from each platform and maintain the users with English language content from before and after July 27, 2022. The final sample is 16,449 users on Gettr, 16,891 on Truth Social, and 9,155 from Gab.

In a robustness exercise, for the Gettr platform, I identify a subsample of US based users who list their location at the county level in their profiles resulting in 14,430 users. I also create a secondary dataset of 14,639 known Brazilian users. These two samples allow me to explore country specific contemporaneous shocks which may confound estimates and assess for platform led moderation of various sensitive events which are geo-specific.

In addition to the raw counts of posts, comments, and replies, I collect metadata on each user's likes per post, repost frequencies, and other engagement behaviours where available. These data

[23] See Thomas Barrabi, "Meta exec's frantic warning about Instagram's alarming 'fake' activity numbers in spotlight at FTC trial," *New York Post*, May 14, 2025. Available at: https://nypost.com/2025/05/14/business/meta-executive-warned-fake-engagement-on-instagram-app-could-be-in-range-of-40-court-docs/.

enrich the analysis by enabling investigation not only of content production but also of patterns of user engagement and amplification across platforms. I omit reporting or including user's follower or following counts in any analysis, as I do not have data on pre-treatment counts of followers and any post treatment following count is surely endogenous to content moderation.

3.2 Measurement of Threatening Content using the Perspective API

To evaluate the extent to which user-generated content contains threatening language, I use Google's Perspective API (Wulczyn et al., 2017; Dixon et al., 2018). The API produces a machine-learning-based prediction for several dimensions of perceived harmfulness. My primary measure is the API's "threat" score, which ranges from 0 (not threatening) to 1 (most threatening). According to the API documentation, threatening language is defined as "an intention to inflict pain, injury, or violence against an individual or group." I provide examples along the continuum of threat scores in Table A.4 for reference.

Each post and comment from Truth Social, Gettr, and Gab is scored consistently using the same API version, classification settings, and post-processing pipeline to ensure comparability across platforms and over time.

In addition to the threat score, the Perspective API returns related scores including toxicity, severe toxicity, identity attack, insult, and profanity.[24] Although the analysis focuses primarily on the threat score, I report robustness checks using these alternative dimensions.

I repeat the exercise on the alternate Brazilian sample using the Portuguese version of perspectives API.

3.3 Measurement of Discourse around Sensitive Events

To examine the prevalence of "sensitive-events" in user content, I focus on five politically salient topics frequently implicated in content moderation debates: 1) January 6 commentary, 2) election fraud claims, 3) anti-vaccine discourse, 4) conspiratorial narratives, and 5) climate change skepticism.[25]

The five topics were selected because they align with the policy's definition of sensitive events

[24] See https://developers.perspectiveapi.com/s/about-the-api-attributes-and-languages?language=en_US for documentation on all available attributes.

[25] Definitions of each category are provided in Table C.2.

and reflect issue areas that have triggered scrutiny by both economic researchers and platform governance bodies (Guess et al., 2020; Berlinski et al., 2023; Aridor et al., 2024).

Content was classified using a rule-based dictionary method. For each topic, I constructed a hand-curated list of regular expressions designed to match distinctive phrases and terms representative of the discourse. The dictionaries draw from domain expertise, prior literature, and manual inspection of platform-specific vernacular. Each post was then flagged with a binary indicator for each topic based on whether any of its associated patterns were detected. This approach enables high-recall classification of topic-specific discourse at scale while maintaining transparency and interpretability in the construction of labels.

I aggregate three of these categories: election fraud, anti-vaccine content, and QAnon-related discourse, into a single composite indicator for conspiratorial narratives:

$$\texttt{conspiracies}_i = \mathbb{1}\{\texttt{election\_fraud}_i = 1 \vee \texttt{anti\_vax}_i = 1 \vee \texttt{qanon}_i = 1\}$$

Sample details and descriptive statistics on the distribution of threat and topics scores from before the policy change are provided for each platform in Table A.3.

4 The Effect of Google's Policy on Platform Content Moderation

This section begins by addressing the identification challenges in estimating the causal effect of Google's policy change on platform content moderation. I then outline the empirical strategy and present within-platform results, robustness checks, and causal estimates.

4.1 Identification

Google's July 27, 2022 Play Store bulletin provides a unique quasi-experimental lever to study whether upstream distributors can compel downstream platforms to moderate speech. At the moment of the announcement, the three largest "Alt-Tech" social platforms occupied markedly different positions in the Android distribution chain: Gettr was already listed; Truth Social had a pending application whose approval now hinged on compliance; and Gab, having been permanently delisted in 2017, remained entirely outside the Play Store. As a result, the announcement triggered a sharp, platform-specific change in regulatory exposure that was orthogonal to the behaviour of any individual user. While Gettr retained its listing and Truth Social was eventually admitted,

whether either platform actually altered its moderation practices remains an empirical question.

To identify the causal effect of the Google policy on moderation, measured via changes in users threatening content, I implement a triple-differences (DDD) design. This approach leverages variation along three dimensions: between users (those with high vs. low pre-policy threat tendencies), over time (before vs. after the policy), and across platforms (those affected by the policy vs. Gab, which was unaffected). This design isolates the policy's effect on users exposed to it through their activity on covered platforms, relative to comparable users on an untreated platform. The validity of this strategy rests on standard DDD assumptions, including parallel trends across treatment groups in the absence of intervention, and no behavioural response prior to the announcement (no anticipation), which I address by way of event study analyses both within and between platforms. In addition, several identification threats complicate interpretation. These include: mechanical threats comprising incomplete exposure to treatment and data loss due to retroactive content takedown; selection and attrition; contemporaneous shocks; user selection into platforms; voluntary user self-censorship; and spillovers across platforms.

Mechanical threats. Although the Google Play Store policy applied specifically to Android apps, user content is accessible across multiple interfaces, including web and iOS, where platform incentives to moderate may differ. If developers implemented content moderation only for Android clients while allowing more permissive content on web or iOS, then threat levels observed through web scraping could understate the true effect. However, such selective moderation is technically complex, resource-intensive, and uncommon in practice, as it requires maintaining divergent infrastructure across platforms. Moreover, platform moderators can retroactively remove policy-violating posts from the global content pool. Since the web crawler captures only surviving posts from the web based version of the platform, the estimated decline in threatening content is a conservative, lower-bound measure of the platform's full moderation response.

Selection and attrition. A key identification concern is non-random user selection around the time of treatment. The main specification includes only users active both before and after the Google policy change, which risks selection into the analysis panel. If users with higher pre-treatment threat scores are more likely to remain, the post-policy sample will overrepresent high-threat users, potentially attenuating moderation effects by limiting scope for observed declines. Conversely, if higher-threat users disproportionately exit before the policy, the remaining sample will underrepresent those most likely to be moderated, biasing estimates upward.

Second, even conditional on inclusion, there may be selective attrition if users with different threat levels exit at different rates after the policy. If higher-threat users disproportionately leave after treatment, the observed decline in threatening content may partly reflect compositional shifts rather than platform enforcement or remaining behavioural change.

To directly assess both threats, Table A.5 presents two sets of results. For each platform, I make use of the full random sample, which includes users who posted before and after the policy shift, as well as users who presented only before. Columns (1)–(6) show that users with higher pre-treatment threat scores are significantly more likely to remain in the post-policy panel on all three platforms. The same is replicated for Gab in Table A.6 This selection is strongest on Gab, moderate on Truth Social, and smallest on Gettr, but consistently positive and significant. Because these users are more likely to post threatening content, their overrepresentation in the post period may exaggerate the estimated moderation effect, biasing the coefficient downward (i.e., making it more negative than the true effect).

Columns (7)–(10) then assess how long users remain active after the policy. On Gettr, higher-threat users tend to exit sooner, suggesting either enforcement targeting or voluntary disengagement. On Truth Social and Gab, by contrast, higher-threat users remain active longer, which might be interpreted as limited banning or greater persistence among the most extreme users. Sample and post-treatment heterogeneity in attrition across platforms implies the need to test whether sample selection and attrition distort the main estimates.

I implement two such robustness checks in Section 4.3.2. First, I re-estimate the main specification using a balanced panel of users observed throughout the entire period (i.e., those present at endline) and then test whether estimates are stable across rolling monthly user exit bins. Second, I compare estimates from the full sample including users not present after treatment to those in the full panel, excluding user fixed effects to retain entrants. In both cases, I find no meaningful differences in estimated treatment effects.

Contemporaneous Shocks. Another central challenge, given the design is the possibility that concurrent events, such as political developments, parallel policy changes (e.g., from Apple), or broader competitive pressures within the platform ecosystem, may have occurred around the same time as Google's announcement, confounding the interpretation of any observed content moderation. In particular, changes in threatening content might reflect a general shift in online discourse rather than a response specific to Google's distribution policy. My empirical strategy addresses this concern in three ways.

First, I use Gab as an untreated counterfactual. Gab shares the same language environment and U.S.-oriented user base as the treated platforms but was entirely excluded from the Google Play Store. Any broad shocks to online speech, such as national political events or shifts in discourse norms, should plausibly affect Gab as well. To validate this assumption, I assess both behavioural trends and linguistic similarity across platforms using pre-treatment event studies and platform-level comparisons. To evaluate content similarity, I extract weekly n-gram frequencies by platform and compute pairwise similarity metrics (Jaccard and cosine) to track convergence or divergence

in language use over time. As shown in Figure B.2, Gab and Gettr exhibit consistently high pre-treatment similarity (cosine $\approx$ 0.5–0.6), while Truth Social remains linguistically distinct from both, likely reflecting differences in platform culture and agenda-setting. Around the policy announcement, both the Gab–Gettr and Gettr–Truth Social pairs show a sharp, short-lived drop in similarity, consistent with a sudden divergence in content likely driven by platform-specific moderation responses. Post-treatment, similarity gradually increases, with Truth Social in particular converging toward Gab and Gettr in the run-up to the 2024 U.S. presidential election, suggesting a re-alignment or normalisation of discourse over time. Taken together, these patterns reinforce the use of Gab as a valid counterfactual and reduce concern that observed changes on the treated platforms are driven by idiosyncratic or coincidental shifts in language or topic salience.

Second, the event-study design allows for a precise assessment of the timing of content changes relative to the Google policy announcement. By estimating dynamic effects across the full study period, the design enables the detection of both anticipatory patterns and delayed responses, while also providing a built-in placebo for events that occur at other points in time. I use this framework to assess three high-salience events that might plausibly influence threatening content: (1) the FBI's August 8 raid on Donald Trump's residence and the August 19 congressional letters sent to platform CEOs; (2) Elon Musk's acquisition of Twitter in late October; and (3) the U.S. midterm elections, which unfolded over several months. In addition, Table C.3 documents Apple App Store policy changes throughout the year. The only potentially relevant update, a revision on October 25 to the "objectionable content" clause, occurred well after the period of interest and closely mirrors Google's earlier policy. This timing is consistent with a leader–follower dynamic in distributor enforcement and is further ruled out by placebo tests showing no discontinuity around the Apple update.

Third, I conduct a cross-country falsification test using a sample of 14,639 Portuguese-speaking users in Brazil on Gettr. These users operate within a distinct political and linguistic environment, largely insulated from U.S.-specific news cycles and cultural cues. If the observed changes were driven by U.S. political shocks or broader shifts in English-language discourse, we would not expect to see similar patterns among Brazilian users. By testing whether any discontinuity occurs around the Google policy announcement within this independent context, the analysis provides a strong check on the language- and country-specific validity of the identification strategy.

Selection into platforms. A further concern is that user selection into platforms may interact with the DDD structure in a way that affects interpretation. In particular, if users with higher pre-treatment threat levels disproportionately sort into treated or untreated platforms, the second and third differences (user-level threat and platform exposure) may not be orthogonal. This would threaten the assumption that differences in threat evolution across platforms can be attributed to treatment rather than underlying user composition. This concern is mitigated by the fact that the

DDD design incorporates user fixed effects, ensuring that within-user changes are isolated. Since the identifying variation comes from how users with different pre-treatment threat intensities adjust over time within each platform, and how these changes differ across platforms, baseline differences in user composition do not mechanically bias the results. Nonetheless, to assess the plausibility of this assumption, I examine the distribution of pre-policy threat scores across platforms. All platforms contain substantial variation in user-level threat scores, with broadly overlapping distributions in both raw and standardized scores (Figure B.3). Supporting this interpretation is the high degree of pre-treatment cosine similarity in linguistic content across platforms (see previous subsection). As an additional check, I replicate the DDD results using matched subsamples with equivalent threat-score distributions across platforms (see robustness checks in Section 4.4).

User self-censorship. Further, attributing any observed change in content to platform moderation requires ruling out the possibility that users adjusted their behaviour in anticipation of increased enforcement, independently of any formal platform response. Distinguishing between platform-initiated and user-initiated changes is difficult without access to internal moderation logs, but behavioural shifts by users are nonetheless a function of platform incentive structures. With this in mind, I expect that user behaviour may gradually adapt in response to enforcement, especially in the longer term, as users internalise new platform constraints. Such adaptation is likely to shape longer-run dynamics and is examined separately in the user behaviour analysis. However, the key threat to identification is anticipatory self-censorship that precedes or coincides with the treatment, which could confound estimates of platform-led moderation. While it is theoretically possible that some users engaged in self-moderation, either voluntarily or in response to implicit signals, this explanation is unlikely for several reasons.

First, users would need to have known that stricter enforcement was forthcoming. This would require either active monitoring of Google's Developer Program policy, exposure to media coverage, or direct platform communication. Yet, a search of mainstream and technology-focused media sources, including blogs, using Google News and newspapers.com yields no evidence of coverage of the July 27 policy update. While a few articles mention contemporaneous changes to app policies, none refer to the "sensitive events" clause or required platform moderation at all. Google Trends data likewise show no spike in public interest in content moderation around the update date, but instead reveal increased search activity only in late October, when Elon Musk's takeover of Twitter sparked broader public debate over content moderation (see Figure B.1).

Second, the platforms in question explicitly marketed themselves as "free speech" alternatives to mainstream platforms, making it unlikely that they would signal to users either explicitly or implicitly that stricter content controls were imminent. While covert enforcement mechanisms like 'shadowbanning' are possible, there is no evidence of public-facing campaigns urging users to moderate their own content. I compare the public moderation policies of Truth Social and Gettr

before and after July 27 using the Internet Archive's Wayback Machine,[26] and find no documented changes in either case. Third, I analyse user-level posts for any rise in discussions around moderation during this period and find no immediate increase in references to moderation or enforcement in user discourse.

I empirically assess the plausibility of self-censorship in several ways. First, the event study design allows me to examine whether content began declining prior to the policy date, which would indicate anticipatory self-moderation. This approach is particularly relevant to the Truth Social case. Since there were no public announcements regarding Truth Social's pre-approval violation of Google policies until August 19, 2022, nor to the best of my knowledge any media coverage prior to this date, I can granularly check for content adjustments which preceded public communication at a granular (i.e. daily) level. If users were responding to platform signals rather than direct enforcement, any shift in content prior to August 19 would be difficult to attribute to user self-censorship, especially considering non of these users are Android users before at least October 22, 2022, when Truth Social was admitted to the Google Play store.

Second, I assess the distribution of changes in content intensity across users, examining whether reductions are concentrated among the most threatening users or distributed over all users. This approach is particularly useful to assess adjustments in the immediate period after the policy announcement. Third, I leverage cross-country variation: if users were moderating in response to Google's policy, effects should be larger in markets where Android dominates. I compare the magnitude and timing of responses between the U.S. (where Android usage is around 45%) and Brazil (where it exceeds 90%).

Spillovers. Lastly, there must be no spillovers across platforms. Moderation activity on Google-exposed platforms should not influence user behaviour on Gab. Spillovers could occur if users migrate to Gab following moderation, or if users active on multiple platforms adjust their content on Gab in response to moderation received elsewhere. I test for spillover utilising a sub sample of users known to be present on both Gettr and Gab.

4.2 Empirical Strategy

My empirical approach begins by focusing within platforms to examine whether there is evidence of moderation following the Google policy signal. Specifically, I assess changes in the level of threatening content posted by users who, in the pre-announcement period, exhibited particularly high levels of such content. I compare content adjustments as a function of users pre-treatment 'threateningness' on each platform before and after Google's announcement. If moderation occurred, either by

[26] https://archive.org/

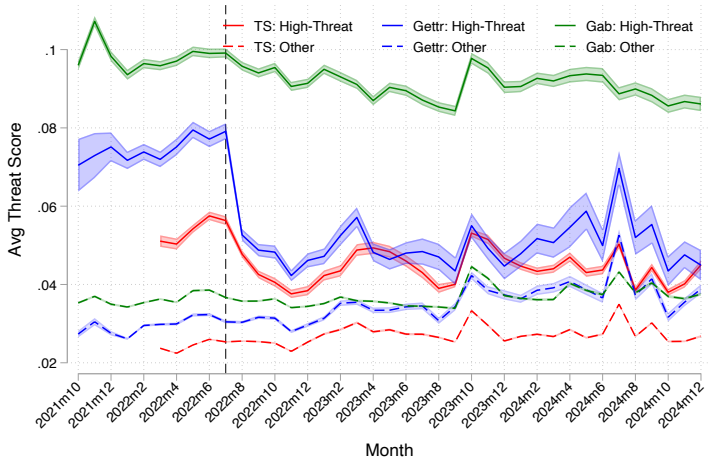
the platform or through user self-censorship, we would expect to see a decline in the average threat level of posts from higher threat users.

I begin by plotting the evolution of average post threat scores, comparing users whose pre-announcement posts were above the 90th percentile of the threat score distribution ("High-Threat") to all other users. Figure 4.1 shows that, prior to the announcement, threatening content on each platform evolved in parallel across user groups. More importantly, similar groups evolved in parallel between platforms.

Following the announcement, however, average threat scores among high-threat users decline sharply on both Gettr and Truth Social after the July 27 Google announcement, continuing to fall for another month before returning to parallel trends with other users on each platform around September 30. Truth Social starts from a lower pre-treatment average and exhibits a more modest decline. It also shows periods of rising threat scores that move in parallel with Gettr, though they never return to pre-treatment levels. Unlike the divergence seen between July and September 2022, these post-treatment movements consistently track those of low-threat users.

In absolute terms, the post-treatment mean for high-threat users converges to a statistically similar level across both Google-exposed platforms, suggesting a common moderation threshold required for compliance. Among non-high-threat users, there is an overall upward trend in threatening content during the same period. Finally, no comparable reduction is observed among high-threat users on Gab. While there is a slight, gradual convergence toward the threat levels of other users, there is no sharp decline at any point in the series. This is the first indication that the changes seen on Gettr and Truth Social are attributable to Google's policy shift.

Figure 4.1: Threat score of users posts across platforms before and after Google announcement



Notes: This figure plots the average threat score of posts across Gettr, Truth Social and Gab. I split users within platform based on the threat score of all of that platforms users posts in the pre-exposure period. "High-Threat" users are those whose posts before the Google announcement were on average above the 90th percentile of the threat score distribution by platform. "Other" are all remaining users. The shaded areas indicate 95% confidence intervals for the means.

To formally test for a shift in threat content within the covered platforms, I implement a difference-in-differences design that exploits variation across users in their pre-announcement threat levels and across time relative to the policy signal. The empirical specification is:

$$Y_{pit} = \beta(Post_t \times \overline{Threat_i}) + \alpha_i + \delta_t + \epsilon_{it} \tag{1}$$

Here, Y_{pit} denotes the threat score of post p by user i on date t; $Post_t$ is an indicator equal to 1 for dates after July 27, 2022 (the public announcement), and 0 otherwise; $\overline{Threat_i}$ is the user's average threat score in the pre-announcement period, standardised within platform to allow effect sizes to be interpreted in standard deviation units and compared across platforms; and α_i and δ_t are user and year-day fixed effects, respectively.

To capture the average treatment effect on the treated (ATT), I extend the design to a triple-differences framework, comparing shifts in threatening content by pre-treatment user threat scores between covered platforms (Gettr and Truth Social) and the uncovered platform (Gab).

$$
\begin{aligned}
Y_{pit} = \beta &\left(Post_t \times \overline{Threat}_i \times Exposed_i\right) \\
&+ \gamma_1 \left(Post_t \times \overline{Threat}_i\right) \\
&+ \gamma_2 \left(Post_t \times Exposed_i\right) \\
&+ \gamma_3 \left(\overline{Threat}_i \times Exposed_i\right) \\
&+ \alpha_i + \delta_t + \epsilon_{it} \qquad (2)
\end{aligned}
$$

Where, $Exposed_i$ is a binary variable equal to 1 for Gettr and Truth Social, and 0 for Gab. In the most restrictive version of this specification I include user-platform fixed effects, and platform-week fixed effects.

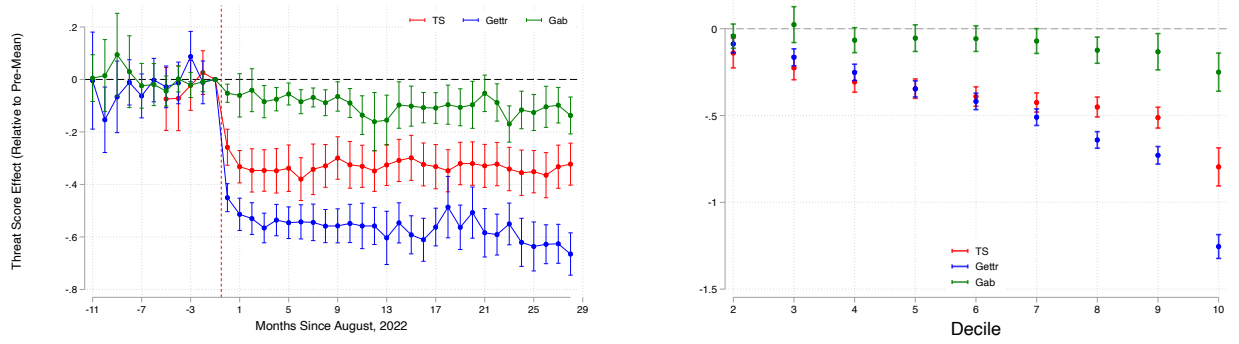
4.3 Results

I begin with a dynamic version of Equation 1, in Figure 4.2a, which replaces the Post indicator variable with dummies for the months around the Google announcement. This specification enables assessment of evidence of within platform parallel trends between users, the timing of any observed changes in competent and the direct comparison between platforms and provides compelling evidence that the observed decline in threatening content on Gettr and Truth Social was a direct response to the Google policy update. The figure indicates that before Google's announcement, users with high and low pre-treatment threat scores followed similar trends in how threatening their posts were across all three platforms. Immediately following the announcement, 'threateningness' dropped sharply and remained consistently below its pre-announcement level for the entire study period on both Truth Social and Gettr as a function of user pre-treatment 'threateningness', but not on Gab. For each one standard deviation in increase in pre-treatment 'threateningness', the likelihood of a post being classified as threatening after July 27 declines by between 50% - 60% on Gettr and 30% - 40% on Truth Social, relative to the pre-treatment mean on each platform respectively. There is a relatively small and marginally significant decline in threatening content on Gab over the longer term but this does not begin until four months after the Google announcement and suggests trend rather than moderation. Importantly, over the entire period, there are no other sharp adjustments following the policy shift, indicating that this was a unique and temporally isolated shock to platform moderation, with no comparable shifts in threatening content occurring later in the study period. Weekly versions of these event studies in Figure B.6 highlight the temporal proximity of the reduction in threatening content to the policy shift on Google exposed platforms. Within one week of the announcement, both platforms see a sharp, statistically significant and persistent decline compared to the week before the announcement. This highlights

how quickly these platforms are able to moderate when incentivised to do so and rules out potential confounders, which occur later than July 27, driving the reduction in covered content.

In Figure B.5a, I estimate a flexible specification that interacts each decile of the pre-treatment threat score distribution with the post-Google-announcement indicator, using users in the lowest decile (bottom 10th percentile) as the reference group. The results show that the effect of the announcement increases monotonically with user 'threateningness' for both Gettr and Truth Social. Relative to users in the lowest decile, users in the 10th percentile (i.e., average pre-treatment content threat scores between [5.5–66] on Gettr and [4–55] on Truth Social) experience a reduction in threat score of 0.043 and 0.022, respectively, corresponding to decreases of approximately 125% and 80% relative to each platform's pre-announcement mean (see Figure B.4 for level estimates). On Gab there is no monotonicity in the effect: there is an economically small and roughly constant, marginally significant decline across deciles until the highest decile, which exhibits a decline of just under 25% of the pre-treatment mean.

Figure 4.2: Effect of Google Announcement on User Threatening Content Across Platforms



(a) Event Studies

(b) Decile Regressions

Notes: Panel a) shows platform-specific regressions of user post threat scores on the interaction of pre-treatment user average standardised threat scores with a set of month indicators relative to Google's announced policy change. The omitted category is July 2022. The dependent variable is scaled relative to each platform's pre-treatment mean and can be interpreted as proportional changes compared to the mean. Panel b) shows platform-specific regressions of the change in users' daily threat scores compared to the platform specific pre-treatment mean on a set of indicators for users' cross-sectional pre-treatment decile of average threat score. The omitted category is the first decile. In both panels, regressions include user control variables interacted with year-day fixed effects, user fixed effects, and year-day fixed effects. Standard errors are clustered at the user level. *** p<0.01, ** p<0.05, * p<0.10.

In Figure B.4b, I estimate the same specification but restrict the post-treatment period to the first two months (i.e., August–September), capturing only the immediate platform response to the Google policy. The results again show a clear monotonic pattern on the Google-exposed platforms, with the magnitude of the effect slightly smaller than in the full-period estimates. In contrast, Gab shows no such monotonicity, and the overall magnitude remains small and flat across deciles.

This short-term pattern is difficult to reconcile with anticipatory self-censorship, which would likely produce a more diffuse reduction across user types, if anything, affecting lower- and mid-threat users more uniformly. Instead, the disproportionate decline concentrated among high-threat

users strongly suggests targeted platform enforcement, consistent with either manual or automated moderation directed at policy-violating content. The absence of a comparable response on Gab further supports the interpretation that this initial moderation response is driven by platform behaviour, not user-led adjustment.

I now turn to the benchmark difference-in-differences estimates from Equation 1. Each column of Table 4.1 presents the estimated effect of a one-standard-deviation increase in a user's pre-announcement threat score on the average 'threateningness' of their visible (i.e., unremoved) content after Google's announcement. The left panel are estimates for Gettr, while the right are for Truth Social. In each panel, column (1) includes user fixed effects, column (2) adds year-day fixed effects, column (3) adds user specific pre-treatment level controls × year-day fixed effects, and column (4) adds a user specific linear trend, which relaxes the assumption that users with different pre-treatment threat average scores follow parallel trends. All specifications indicate a significant reduction in the 'threateningness' of posts on Google exposed platforms after the July 27 announcement. The results are very stable over all specifications even with the inclusion of user-specific linear trends. The estimates imply that the announcement was associated with a reduction in average threat scores of approximately 52% on Gettr and 32% on Truth Social for users who were one standard deviation above the mean in pre-announcement threat levels, consistent with increased moderation targeting high-risk users in response to the policy change.

Estimates for Gab in Table 4.2 highlight a small (<4% compared to pre-treatment mean) and significant overall average decline in threatening content on Gab as a function of user 'threateningness' after the Google announcement. Results are noticeably smaller than other estimates and less robust to the inclusion of user trends than Google distributed app estimates.

Table 4.1: Google Announcement and Online User Threatening Content

	Gettr				**Truth Social**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{Threat}$	-0.0184***	-0.0181***	-0.0179***	-0.0174***	-0.0084***	-0.0082***	-0.0083***	-0.0083***
× Post	(0.0005)	(0.0005)	(0.0004)	(0.0006)	(0.0006)	(0.0006)	(0.0006)	(0.0006)
N	2,968,879	2,968,879	2,968,879	2,968,879	11466215	11466215	11466215	11466215
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
User Control × Time FE	No	No	Yes	Yes	No	No	Yes	Yes
User Time Trend	No	No	No	Yes	No	No	No	Yes
# of Users	25952	25952	25952	25952	13515	13515	13515	13515
DV Pre-Mean	0.0336	0.0336	0.0336	0.0336	0.0265	0.0265	0.0265	0.0265
Treatment SD	0.0819	0.0819	0.0819	0.0819	0.0657	0.0657	0.0657	0.0657

Note: The table presents the results of estimating Equation 1. The dependent variable is the threat score of the users posts on platforms Gettr and Truth Social. $\overline{Threat}_i$ is the pre-treatment average threat score of user i's posts. User post threat score is standardised within platform. $Post_t$ is an indicator variable = 1 following Google's announcement on July 27, 2022, and 0 otherwise. The level of observation is user × year-day. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table 4.2: Google Announcement and Online User Threatening Content (Gab)

	Gab			
	(1)	(2)	(3)	(4)
$\overline{Threat}$	-0.0036***	-0.0035***	-0.0035***	-0.0018**
× Post	(0.0011)	(0.0011)	(0.0011)	(0.0009)
N	13204938	13204938	13204938	13204938
User FE	Yes	Yes	Yes	Yes
Year-Day FE	No	Yes	Yes	Yes
User Controls × Time FE	No	No	Yes	Yes
User Time Trend	No	No	No	Yes
# of Users	7374	7374	7374	7374
DV Pre-Mean	0.0448	0.0448	0.0448	0.0448
Treatment SD	0.0979	0.0979	0.0979	0.0979

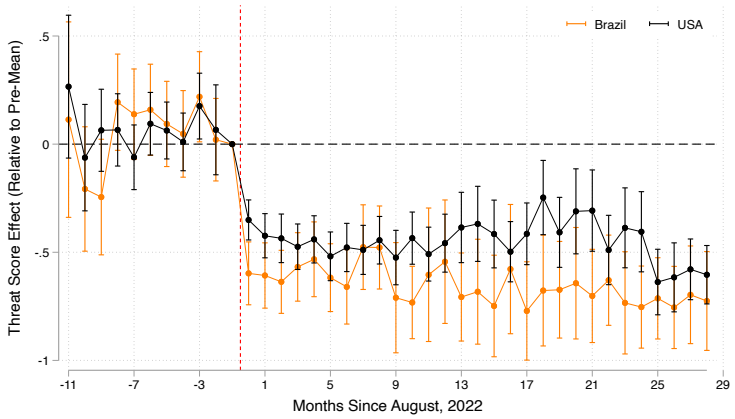
Note: The table presents the results of estimating Equation 1. The dependent variable is the threat score of the users posts on Gab. $\overline{Threat}_i$ is the pre-treatment average threat score of user i's posts. User post threat score is standardised within platform. $Post_t$ is an indicator variable = 1 following Google's announcement on July 27, 2022, and 0 otherwise. The level of observation is user × year-day. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

4.3.1 Cross-Country Falsification: Brazil versus the United States

A central identification threat is that some contemporaneous, U.S.-specific shock, for example, the FBI search of Donald Trump's Mar-a-Lago home on August 8, 2022, rather than the Google bulletin, drove the post–July 27 decline in threatening content. A related concern is that the observed effect could reflect Android-only self-censorship: if U.S. Android users independently toned down their rhetoric while iOS and desktop users did not, the average drop might be a compositional artefact rather than true moderation. The fact that we observe any sharp discontinuity suggests that moderation occurs server-side rather than within a specific channel build (e.g., Android only), but it is still possible that some Android users responded independently to the policy shift.

Event study estimates in Figure 4.3 compare threatening content between the U.S. and Brazilian samples. The results highlight three key points: (1) the estimates for the U.S. sample are similar in both magnitude and timing to those in the full English-speaking sample, suggesting that the main results are unlikely to be driven by U.S.-specific dynamics; (2) this interpretation is reinforced by the Brazilian estimates, which track closely with the U.S. series and thus rule out contemporaneous U.S. political or social shocks as primary drivers; and (3) Brazilian users, who disproportionately access these platforms via Android, exhibit no statistically significant divergence from U.S. users, which suggests an Android-specific self-censorship mechanism is unlikely.

Figure 4.3: Effect of Google Announcement on User Threatening Content on Gettr Across Countries



Notes: This figure plots regressions of country specific user post threat scores on Gettr on the interaction of pre-treatment user average threat scores with a set of month indicators relative to Google's announced policy change. The omitted category is July 2022. The dependent variable is scaled relative to each countries pre-treatment mean and can be interpreted as proportional changes compared to the mean. The bars indicate 95% confidence intervals.

4.3.2 Robustness

I conduct a range of robustness checks to assess the credibility of the benchmark results. Each set of tests examines a distinct modelling or measurement choice to ensure the estimated effect is not an artefact of a specific specification.

In Appendix Table A.7, I re-estimate the main specification from Equation 1 with alternative codings of the dependent and treatment variables. Column (1) uses the log of the threat score rather than the level to account for skewness. Columns (2)–(4) replace the continuous measure of pre-treatment 'threateningness' with categorical bins: quartiles in column (2), a top quartile dummy in column (3), and a top decile dummy in column (4). These tests assess whether the effect is concentrated among the most extreme users or more broadly distributed. Across all variants, I continue to find a significant post-policy decline in threatening content on platforms exposed to Google's enforcement. Appendix Table A.8 probes the sensitivity of results to different thresholds for defining high-threat users. Instead of quartiles or deciles, I use a range of cutoff values based on pre-policy average threat scores, including more stringent definitions of low-threat control groups. The negative post-policy shift remains robust across all cutoffs. Appendix Table A.9 tests whether results hold when exposure is defined using a user's maximum pre-policy threat score, rather than their average. I implement this both as a continuous variable and through dummies at various thresholds (e.g., top 10%, top 20%). This alternative characterisation of user "extremity" yields patterns consistent with the baseline, including strong effects even when the control group is restricted to users with very low threat scores.

In Appendix Table A.10, I assess whether effects vary by content type. I separately estimate treatment effects for original posts (column 1), replies (column 2), and reposts (column 3). The largest declines are observed for original posts, followed by replies, with smaller but still significant effects for reposts.

Next, Appendix Table A.11 examines whether the observed effects extend to other dimensions of harmful speech using alternative scores from the Perspective API, specifically, toxicity, severe toxicity, identity attack, insult, and profanity. On Gettr, I observe consistent and significant declines across all measures. On Truth Social, results are similarly negative for most categories, though effects for severe toxicity and profanity are smaller and statistically insignificant, likely reflecting the platform's pre-existing restrictions on such content. Finally, I examine an independent hate speech measure derived from a fine-tuned BERT model (Aluru et al., 2020) in Appendix Figure B.5. I again find a sharp drop in hate speech beginning after July 27 on the two platforms exposed to Google's policy, with no comparable change on Gab.

Lastly, to assess the consequences of selective attrition and panel sample selection for the estimated treatment effects, I conduct three sets of robustness checks. First, reported in Table A.12, columns (1)–(6) test whether differential attrition after treatment biases the results. These specifications splits the samples into i) users who remain active through the end of the study period, eliminating variation in exposure duration, and ii) users who exit before the last month of the sample period in December 2024, capturing those with shorter observed engagement. I compare estimates from the full sample (columns 1 and 4) to those from the balanced "endline" panel (columns 2 and 5) and those from the attrited sample (columns 3 and 6). I find no meaningful differences in the post-treatment interaction effect, suggesting that differential dropout after treatment does not drive the main results.

In addition, in Figure B.7, I plot the estimated post-treatment interaction effect separately for users exiting before each quarterly cutoff, as well as within non-overlapping quarterly exit cohorts. The top row shows estimates for cumulative exit samples, while the bottom row reports results for disjoint quarterly exit buckets. For both Truth Social and Gettr, the treatment effects remain consistently negative and relatively stable across cohorts. In contrast, the estimates for Gab are smaller in magnitude and either marginally significant or statistically indistinguishable from zero. The absence of clear temporal divergence in any of the platforms, and the stability of the effect for those subject to Google's policy, provides further support that selective user attrition does not account for the main findings.

Columns (7)–(10) of Table A.12 then test for selection into the panel, relaxing the requirement that users be active both before and after the policy change. These specifications omit user fixed effects to accommodate those exiting before treatment and compare average treatment effects in the full

sample (columns 7 and 9) to those in the subsample of "stayers", who are active users on both sides of the treatment (columns 8 and 10). Again, the estimated effects are similar in size and precision, indicating that the results are not being driven by selection into the panel based on pre-treatment threat. Taken together, these findings suggest that the observed decline in threatening content is not primarily due to compositional changes in the user base.

To benchmark these results, I replicate all robustness tests on Gab (Appendix Table A.13 through Table A.17), where the Google policy does not apply. Across specifications, the estimated effects are consistently small, statistically insignificant, or weakly signed, reinforcing the interpretation that the observed shifts on Gettr and Truth Social were driven by policy exposure rather than broader trends. Assessment of the consequences of selective attrition and panel sample selection in Table A.18 suggest no selection or attrition bias on Gab.

4.3.3 Other External Events

The event studies provide strong evidence that Google's July 27, 2022 policy announcement triggered a platform-specific reduction in threatening content on Truth Social and Gettr. The patterns are sharp, temporally aligned with the announcement, and absent on Gab, supporting both the parallel trends assumption and the exclusion of major confounders. Nevertheless, before continuing to the causal estimation, it is worth considering whether any external events could have selectively influenced threatening content on the Google-exposed platforms but not Gab. For an alternative explanation to threaten identification, it would need to (1) disproportionately affect Gettr and Truth Social, (2) begin precisely around July 27, (3) induce a discontinuity in threat scores among high-risk users, independent of the Google policy, and (4) affect users in English speaking countries and Brazil equally. I discuss some possibilities and explain why none plausibly account for the results.

Twitter Takeover. Elon Musk's acquisition of Twitter in late October 2022 has been cited as a potential driver of migration from Alt-Tech platforms, possibly lowering threat levels via selective user exit.[27][28] However, the timing is inconsistent with the results. The largest and only discontinuity in threatening content appears in late July, not October. Daily and weekly threat scores show no structural breaks around the Twitter acquisition, and Gab again shows no change, ruling out a sector-wide shift. If extreme users had exited en masse following the acquisition,

[27]Kate Conger and Ryan Mac, "How Elon Musk Changed the Meaning of Twitter for Users," New York Times, October 27, 2023, https://www.nytimes.com/2023/10/27/technology/elon-musk-twitter-year.html.

[28]Sheila Dang, "Twitter is losing its most active users, internal documents show," Reuters, October 26, 2022, https://www.reuters.com/technology/exclusive-where-did-tweeters-go-twitter-is-losing-its-most-active-users-internal-2022-10-25/.

a second drop would be visible in October, yet none appears. This is further supported by the robustness of main results to specifications which test the effects of selective attrition in the previous section.

U.S. Political Events: Mar-a-Lago Raid. One potential confound is the FBI search of Donald Trump's residence on August 8, 2022, which triggered a surge in threatening rhetoric and prompted a congressional inquiry into platform moderation on August 19.[29] However, several facts make this explanation implausible. Most importantly, results from the Portuguese-language Brazil sample show a nearly identical drop in threatening content beginning on July 27, despite the event being neither politically or linguistically relevant in that context. Topic modelling confirms that the event was salient in user discourse on all three platforms in English but not the Brazilian context on Gettr. Moreover, the timing is inconsistent: the observed decline begins precisely on July 27 and is largely complete by mid-August, ruling out the possibility of delayed effects from the Mar-a-Lago raid.

4.4 Causal Estimates

Having established economically large, consistent and robust estimates for within firm reduction in 'threateningness' of visible content, and having ruled out mechanical and contemporaneous threats to identification, I now turn to the causal estimates for Google's policy on platform side moderation of user content. I begin by assessing evidence of third difference parallel trends and ruling out spillovers onto the counterfactual platform Gab. I then turn to estimation and robustness.

To assess the assumption of parallel trends and any anticipation effects, I provide dynamic DDD estimates in Figure B.8. The first thing to note is that there is no statistically significant difference in high and low-threat users between platforms before the policy shift. While there is a slight upward pre-trend, its direction suggests that the estimated ATT may represent a conservative (lower-bound) estimate of the policy's effect on threatening content. Secondly, there is a statistically significant difference in the adjustment in platform level exposure (user pre-treatment threat score) between the Google based apps and Gab beginning at the time of the Google policy shift and persistent over the entire period.

I directly test the plausibility of the no-spillovers assumption by examining two groups: a subsample of 2,555 users active on both Gab and Gettr before the policy shift[30] and users who join Gab shortly after treatment. Column (1) of Table A.23 estimates the effect of the policy on Gab

[29]Drew Harwell and Cat Zakrzewski, "Lawmakers Demand Data about Online Threats against Law Enforcement," *Washington Post*, August 19, 2022. See Figure C.4 for a copy of the letter sent to CEOs.

[30]Users are identified as being on both platforms by the presence of identical names across platforms, a method practised in **?**, Jones et al. (2017), and Rizzi (2024).

users who are not known to be present on Gettr, and finds a small but statistically significant post-treatment decline in threat scores. Column (2) restricts the sample to users known to be active on both platforms and yields a slightly larger significant reduction. In Column (3), I formally test for heterogeneity in this effect by including an interaction for cross-platform users; the coefficient is statistically indistinguishable from zero, indicating no significant difference in post-treatment reductions between the two groups. In columns (4) and (5) I restrict the sample to users present on both Gettr and Gab and tests whether changes in Gab content correlate with these users' pre-treatment threat levels on Gettr. If the moderation on Gettr affects their behaviour on Gab, then this will influence the interpretation of Gab as the counterfactual. In column (4), I interact the users pre-treatment Gettr threat average with the post treatment dummy and find no significant interaction effect. In column (5) I include a third interaction, whether the Gettr user exited Gettr over the prior or not and I, again, find no statistically significant effects. These results suggest that users exposed to Google's policy and active on Gab, did not behave any differently from other existing Gab users on the Gab platform after the policy.

Second, I assess differences in posting behaviour between users who joined Gab after the policy shift and were active on Gettr versus those who joined Gab after the policy but were not known to be active on Gettr. While the primary analysis focuses on users active both before and after the policy change, new joiners may still affect the behaviour of incumbent Gab users through interaction, potentially biasing the estimates. Column (6) of Table A.23 presents average differences between these two groups, showing no statistically significant difference in threat content. In Column (7), I interact group membership with the number of months since the policy change, including time-to-policy fixed effects. I find that Gettr users who joined Gab immediately after the policy shift exhibited significantly lower threat scores than other joiners of the same month, suggesting a potential short-term spillover of lower threatening content across platforms. This effect diminishes over time. This could be interpreted as an influx of users who naturally sorted towards Gettr and have a preference for less threatening content more consistent with the Gettr user base than Gab, but then shift following distaste for moderation. Another interpretation may be that users exposed to moderation of their threatening content on Gettr are less likely to post threatening content on Gab in the immediate aftermath of the experienced moderation. I explore these dyanmics in more detail in Section 4.6. While new Gab joiners with prior Gettr activity initially posted less threatening content, this does not reflect a baseline difference between Gab and Gettr users. As shown in Column (3), cross-platform users active pre-treatment on Gab do not exhibit significantly different pre-to-post trends in threat content. Thus, the temporary difference observed among new joiners in Column (6) likely reflects short-term behavioural spillovers rather than permanent selection. Because the main estimates rely on users active before the policy shift, the validity of Gab as a counterfactual is not compromised.

Table 4.3 provides the triple difference specification, capturing variation across time, users, and

platform-level exposure to Google's policy. The results show that, relative to Gab, threat scores declined significantly more on platforms exposed to Google's enforcement incentives. For each one standard deviation increase in a user's pre-policy 'threateningness', average post-policy threat scores fell by approximately 44 percent more on Gettr and 26 percent more on Truth Social compared to Gab, corresponding to an average reduction of 32 percent across the treated platforms. These differences represent substantial relative changes in content conditional on prior risk level, and point to a clear discontinuity in moderation response linked to infrastructure pressure. The preferred specification includes user-platform fixed effects, which control for time-invariant differences in baseline threat levels and platform usage patterns across users, and platform-week fixed effects, which absorb common shocks and policy changes affecting all users on a given platform in a given week. Results are remarkably stable across all specifications.

Robustness. I subject the triple difference specification to the same set of robustness checks applied in the platform-specific models, adapting them to the three-way interaction. As detailed in Table A.19 through Table A.22, the results remain consistent across alternative functional forms of the pre-treatment threat measure, different definitions of high-risk users, restrictions by post type, and across alternative toxicity measures from the Perspective API. Across all specifications, the estimated differences between treated platforms and Gab remain large, statistically significant, and directionally stable. In another specification, I repeat the exercise in a difference in differences specification excluding the sample to just the top 90th percentile users from each platform, and test the difference between platforms, after the announcement. Results in Figure B.9 show the same sharp reductions in Google based apps at the announcement date. Finally, I replicate the triple difference analysis on a matched sample constructed by aligning the distribution of pre-treatment threat scores across platforms using coarsened bins; results in Figure B.10 confirm that the estimated treatment effects remain consistent.

Table 4.3: DDD - Google Platforms vs Gab Users Threat Scores After Google Announcement

	Platforms				Google			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{Threat}$	-0.0529***	-0.0529***	-0.0522***	-0.0523***	-0.0529***	-0.0529***	-0.0522***	-0.0523***
× Post	(0.0157)	(0.0157)	(0.0155)	(0.0156)	(0.0157)	(0.0157)	(0.0155)	(0.0156)
Gettr ×	-0.4423***	-0.4423***	-0.4424***	-0.4411***				
$\overline{Threat}$ × Post	(0.0227)	(0.0227)	(0.0226)	(0.0225)				
TS ×	-0.2564***	-0.2564***	-0.2557***	-0.2547***				
$\overline{Threat}$ × Post	(0.0299)	(0.0299)	(0.0298)	(0.0298)				
Google ×					-0.3164***	-0.3164***	-0.3159***	-0.3145***
$\overline{Threat}$ × Post					(0.0236)	(0.0236)	(0.0235)	(0.0235)
N	25217118	25217118	25217118	25217118	25217118	25217118	25217118	25217118
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	No	Yes	Yes	Yes	No	Yes	Yes	Yes
User Controls × Time FE	No	No	Yes	Yes	No	No	Yes	Yes
Platform × Year-Week FE	No	No	No	Yes	No	No	No	Yes
# of Users	46106	46106	46106	46106	46106	46106	46106	46106
DV Pre-Mean	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000	1.0000
Treatment SD	2.3541	2.3541	2.3541	2.3541	2.3541	2.3541	2.3541	2.3541

Note: The table presents the results of estimating Equation 1. The dependent variable is the threat score of the users posts on Truth Social. $\overline{Threat}_i$ is the pre-treatment average threat score of user i's posts. User post threat score is standardised. $Post_t$ is an indicator variable = 1 following Google's announcement on August 19, 2022, and 0 otherwise. The level of observation is user × year-day. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

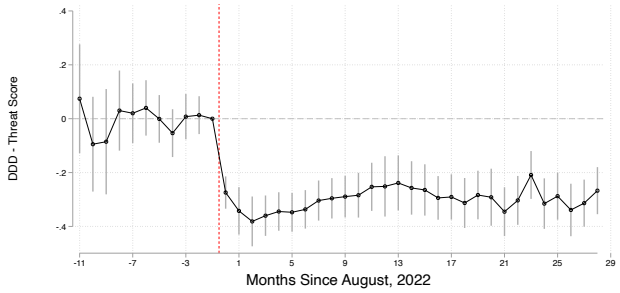
4.5 Topic Specific Responses to Infrastructure Moderation

While Google's 2022 Play Store policy formally targeted violent or inciting content, its revised language extended enforcement to a broader class of "sensitive events," defined as those with "significant social, cultural, or political impact", including civil emergencies, public health crises, and political conflicts. This expansion created the possibility that platforms would not only suppress explicitly threatening speech but also reduce the visibility of topics perceived to carry reputational or regulatory risk, even when such topics were not inherently unlawful or false. To investigate this possibility, I examine how enforcement pressure shaped the prevalence of user posts on five politically salient topics frequently implicated in content moderation debates.

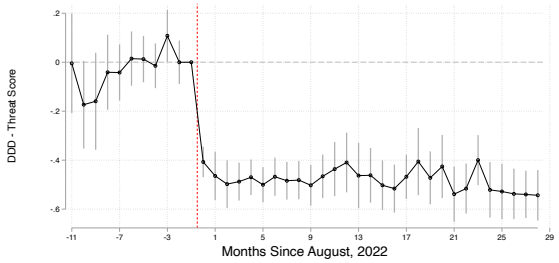
Empirical Approach. I replicate the design introduced in earlier sections and construct a user level measure of pre treatment topic intensity, defined as the proportion of posts made by a user on a given topic during the six months preceding the policy announcement. This proportion is standardised to facilitate interpretation. I then estimate dynamic difference in differences specifi-

Figure 4.4: DDD Threat Google apps vs Gab

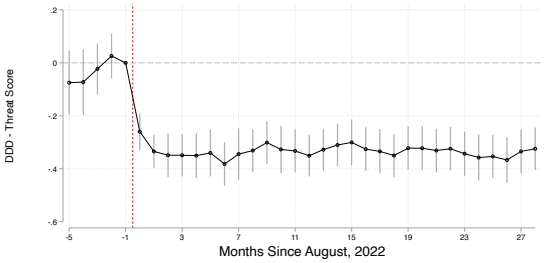
(a) Google vs Gab



(b) Gettr vs Gab



(c) TS vs Gab



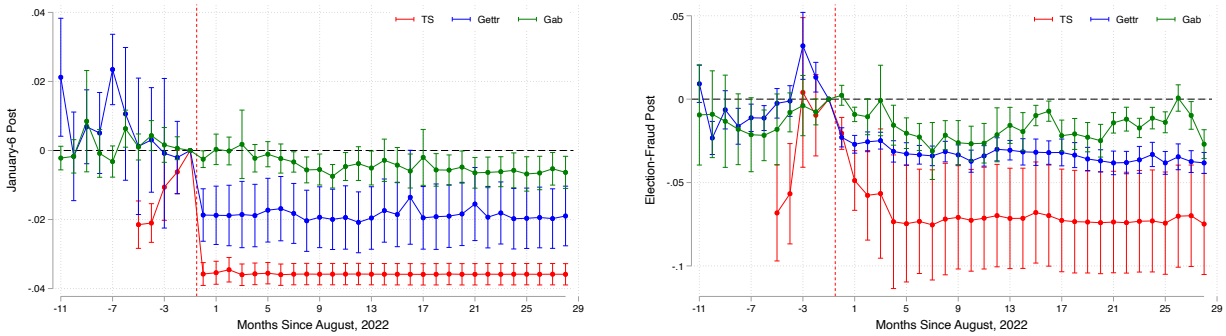
Note: Each panel reports dynamic triple-differences (DDD) estimates from regressions of post-level threat scores on the interaction between a user's pre-treatment average standardized threat score, month indicators relative to Google's policy announcement in late July 2022, and platform indicators. The omitted category is July 2022. This captures differential changes in threatening content on Google-exposed platforms (Gettr and Truth Social) relative to the untreated platform (Gab). Panel (a) aggregates both Gettr and Truth Social into a single treatment group, while panels (b) and (c) report disaggregated estimates for Gettr and Truth Social, respectively. All models include fixed effects for users, platforms, platform-by-month, and user-by-platform combinations. The sample includes posts from August 2021 to January 2025. Estimates are scaled relative to platform-specific pre-treatment means. Standard errors are clustered at the user level.

cations, where the outcome is a binary indicator for whether a user posts on the topic in a given month, and the treatment variable is the interaction between the standardised pre treatment share and time relative to the policy change.

This approach allows me to examine whether users with greater prior engagement on a topic reduced such engagement differentially following the introduction of Google's policy.

Results. The results in Figure 4.6 and dynamic specification results in Figure 4.5 and Figure B.11 reveal a consistent pattern. Among platforms exposed to Google's enforcement regime, users with higher baseline engagement in a topic were significantly less likely to post about that topic after the policy announcement. The effects emerge immediately in August 2022 and persist throughout the post period.

Figure 4.5: Topic-specific event study estimates by pre-treatment user topic average posts proportion.



(a) January 6 commentary

(b) Election fraud claims

Note: Each panel shows the interaction effect between user-level pre-policy topic engagement (standardised) and time relative to Google's July 2022 policy update, using the dynamic difference-in-differences specification from Equation 1. Bars are 95% confidence intervals. Models include user fixed effects and platform-by-month trends. Standard errors clustered at the user level.

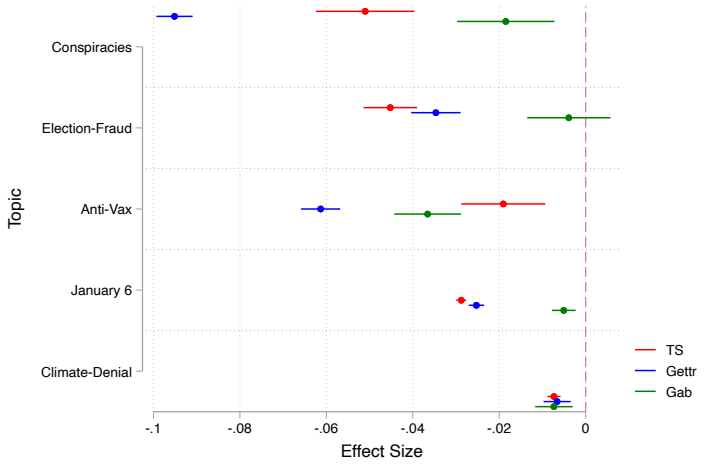
The largest and most sustained effects appear for January 6 commentary, election fraud claims, and conspiracy theories. On both Truth Social and Gettr, users with high prior engagement reduced topic related posting sharply in the first month following the policy change. The magnitude of the decline is greater on Truth Social for January 6 and election fraud[31], and greater on Gettr for anti-vaccine discourse and conspiracy theories.

By contrast, Gab displays a slightly downward trajectory in topic content, similar to the pattern noticed in threatening content. There is no statistically significant sharp reduction in the first few months after the policy change on any topic. With the possible exception of climate change

[31]This is not surprising given the close ties of Truth Social to Donald Trump, and his connections to both topics.

skepticism, where there is a delayed decline around eight months after the announcement, topic related engagement on Gab follows a relatively stable or gradually declining trend. This divergence strengthens the interpretation that observed shifts on Truth Social and Gettr were induced by infrastructure exposure, rather than endogenous user behaviour or platform wide drift.

Figure 4.6: Comparison of salient 'Sensetive Events' discourse across platforms after Google policy announcement



Notes: This figure plots estimated coefficients from user-by-date fixed effects regressions examining the interaction between a post-policy indicator for August 2022 onwards and the user's pre-policy average topic engagement, standardised within each platform. Each panel corresponds to a distinct topic: Election-Fraud, Anti-Vax, Climate-Denial, January-6, and a composite "Conspiracies" variable (election fraud, anti-vax, and QAnon). The dependent variable is a binary indicator for topic presence in a post, and the key independent variable is the interaction term capturing whether more engaged users shifted content more strongly post-policy. Estimates are shown separately for Truth Social (TS, red), Gettr (blue), and Gab (green), with 95% confidence intervals clustered at the author level.

Falsification. As a falsification test, I assess whether users' pre-treatment levels of threatening speech can account for the decline in topic-related posting observed in the main analysis. This exercise evaluates whether platform responses reflect a broader suppression of high-risk users, irrespective of content, rather than targeted enforcement aligned with the specific issue areas referenced in Google's policy. To implement this test, I repeat the dynamic difference in differences specification from Section 4.5, replacing the topic-specific exposure measure with each user's standardised pre-policy share of threatening posts. The outcomes remain unchanged: monthly binary indicators for whether a user posted on each of the five topical categories. The estimates in Figure B.12 reveal no significant association between baseline 'threateningness' and changes in topic-specific posting behaviour. Across all topics and platforms, interaction effects are consistently small and statistically indistinguishable from zero, suggesting that the observed declines in topic engagement are not driven by general enforcement against high-risk users but by issue-specific responses tied

to content exposure.

4.6 Spillovers and Behavioural Responses

The previous sections establish that threatening and sensitive content declined substantially on platforms exposed to Google's enforcement policy. While Section 4.4 presents evidence that spillovers are unlikely to bias the estimated treatment effects, this section investigates whether the policy generated behavioural responses or unintended side effects. Specifically, I assess three channels: (1) declines in posting activity and user visibility; (2) changes in user discourse surrounding moderation prior to exit; and (3) behavioural shifts on Gab among users who exited Gettr, as a test of cross-platform spillovers. The findings suggest that the policy reduced harmful content through a combination of platform enforcement and user disengagement, without triggering backlash, coordinated migration, or reallocation of threatening discourse to less regulated spaces.

To assess user engagement on platforms, I assess exit rates across platforms by pre-treatment threateningness quartiles. I then construct a panel version of the dataset at the user-day level for each platform and repeat the dynamic version of Equation 1, replacing the outcome variable with: i) an indicator for whether a user posted original content on a given day, and ii) the total number of posts a user made on a given day. To measure 'virality', I replace the outcome variable with i) each posts total like count and ii) the total number of likes a user received on a given day.

As exhibited in Figure B.13, the largest proportion of exiters on Gettr occurs in the first few months following the policy shift, with the effect slightly more conentrated above the first quartile of pre-treatement 'threateningess'. There is a modest increase in user exit on Truth Social in the months following the policy, although this effect is overshadowed by a much larger spike in exits during the final quarter of 2023. I interpret this latter decline in users as a response to disclosures of substantial losses in November 2023, an interprtetiaon supported by a second wave of user disengagement following first-quarter reports of over $327 million in losses and just $770,000 in revenue.[32][33][34] There is a gradual incline in the share of disengagement on Gab over the entire period with no sharp rise around July 2022. Overall, the evidence points to a short-term rise in exits on treated platforms following the policy, with longer-run patterns shaped by platform-specific events.

[32]Company performance driven user exit may be explained by declining user confidence in the platform's viability, reduced incentives for marketers or influencers to remain, negative press coverage affecting reputation-sensitive users, and/or anticipatory disengagement by low-activity accounts.

[33]Weprin, Alex (November 13, 2023). "Trump's Truth Social Lost Tens of Millions Since Launch, New Filing Shows." *The Hollywood Reporter*.

[34]Goldstein, Matthew (May 20, 2024). "Trump Media Reports Lackluster Revenues and Large Losses." *The New York Times*.

Results in Figure B.14 show similarly distinct patterns of behavioural change across platforms following the policy shift. On Gettr, users become significantly less likely to post on a given day and reduce overall post volume by roughly 50% relative to the pre-treatment platform average in the long run, though there is some indication of pre-existing trends. Engagement, as measured by likes, also declines steadily among higher-threat users throughout the post-treatment period. This suggests that the policy may have disproportionately reduced the visibility and engagement of threatening users, either through demotion or self-censorship. On Truth Social, there is an initial 10% decline in posting following the policy announcement, which gradually reverts to pre-treatment levels over the subsequent 12 months. Likes remain stable in the short run but increase over time, particularly among high-threat users, in line with a resurgence in posting. In contrast, Gab exhibits a steady decline in posting among higher-threat users, even before the policy shift, while likes remain relatively flat for several months before entering a gradual downward trend. This could be interprerteded as endogenous disengagement or shifting user dynamics, but the post-policy dip in likes may also indicate latent spillover effects or audience fragmentation as the ecosystem adjusts.

While these trends show aggregate behavioural changes, they do not reveal whether users perceived or reacted to moderation pressure in ways that prompted exit or disengagement. To explore this possibility, I next examine user discourse in the days surrounding exit. Figure B.15 examines whether high-threat users became more likely to discuss moderation in the period immediately preceding exit and whether this pattern changed after the policy. I identify all users who exited in the three months before and after the July 2022 policy announcement on Gettr and flag whether they mentioned moderation in the 30 days prior to their last post. I then regress this outcome on pre-treatment threat deciles, separately by platform and period. Prior to the policy, all three platforms exhibit a similar pattern: moderation discourse rises with threat decile but levels off in the upper range. This suggests a baseline tendency for more threatening users to express concerns about moderation before exit. In the three months after the policy, however, platform trajectories diverge. On Gettr, there is a sharp and monotonic increase: users in the highest deciles are 13–15 percentage points more likely than low-threat users to mention moderation before leaving. On TS, this pattern is weaker and only marginally significant in the top decile. Gab shows no such relationship. When the post-treatment window is extended to include all users exiting through October 2024, the pattern on Gettr and TS remains strong and significant, while the effects on Gab flatten further. These findings suggest that only platforms exposed to Google's enforcement pressure exhibited a post-policy rise in threat-linked moderation discourse preceding user exit.

Finally, to assess whether exiting users transferred harmful content to less regulated platforms, I examine changes in their Gab behaviour around the time of their Gettr exit. While Table A.23 already shows no significant shift in Gab content among cross-platform users after the policy, this analysis focuses on individual-level dynamics in the days surrounding exit. I estimate separate event

study models for users who exited Gettr before and after the policy, interacting event time with policy-period status and controlling for user and date fixed effects. As shown in Figure B.16, users who exited after the policy display a short-lived, statistically significant increase in threatening content on Gab just before exit, followed by an immediate return to baseline. This pattern is not observed among pre-policy exiters. I also find no post-exit increase in moderation-related discourse, nor any divergence between groups. These results provide further evidence that the policy did not trigger the kind of user migration or content reallocation observed in earlier studies of public deplatforming.

Taken together, the results suggest that Google's policy reduced harmful content through a combination of platform enforcement and selective user disengagement, particularly among more threatening users on Gettr. I observe a clear short-term rise in user exit following the policy, but this exit was not accompanied by a measurable increase in threatening content or moderation-related discourse on Gab. Nor do I find evidence of backlash or coordination typical of previous deplatforming episodes. Instead, the response appears fragmented and individualized, marked by reduced posting, lower engagement, and exit from treated platforms without detectable reallocation of harmful discourse to alternative sites. While I cannot observe all potential destinations, the evidence within the observable ecosystem points to reduced activity rather than strategic migration.

One possible explanation for this muted response lies in the nature of enforcement. Unlike prior episodes involving coordinated bans or high-profile moderation crackdowns (Agarwal et al., 2022; Horta Ribeiro et al., 2023; Rizzi, 2024), the Google policy relied on infrastructure-level levers, quietly conditioning app store access rather than directly removing users. This relatively silent enforcement may have limited public backlash, reducing the visibility and ideological salience that often catalyzes mass user exits. Moreover, by restricting availability rather than enacting outright bans, the policy may have encouraged gradual disengagement rather than coordinated migration. In this sense, infrastructure-led governance can constrain harmful content without provoking the redistributive dynamics seen in prior moderation campaigns.

5 Policy Implications

My findings reveal how dominant app distributors, those that control a key access point between users and platforms, can act as de facto regulators of content moderation. Through their control of mobile distribution infrastructure, these firms can compel platforms to adjust their moderation practices, not by directly managing platform content, but by conditioning continued access to users on compliance with content standards. This mode of private governance has far-reaching implications for the regulation of digital markets and online speech.

First, market power in distribution can translate into governance power over information flow, even in the absence of formal regulatory authority. When access to users is bottlenecked through a single intermediary, that intermediary gains leverage to impose content standards with potentially global effects. This raises fundamental concerns about the legitimacy of private governance, particularly when such enforcement lacks transparency or democratic oversight. At the same time, there is growing evidence that moderating online content can yield real-world benefits: hate speech online has been shown to increase offline hate crimes (Müller and Schwarz, 2021), while content moderation can reduce such violence (Duran et al., 2022). The exercise of private governance thus poses a tradeoff, between concentrated, opaque enforcement and potentially substantial public benefits.

Second, enforcement by a single distributor can produce system-wide effects. When a platform modifies its moderation policies in response to distributor pressure, these changes often apply across the platform, affecting all users; not just those accessing the service through the distributor's channel. As a result, enforcement by one actor can reshape behaviour across the broader digital ecosystem.

Third, this dynamic creates the potential for free-riding among other distributors. Platforms may incur the costs of compliance to retain access to a dominant distribution channel, while users accessing the same platform through other means, such as web or alternative app stores, benefit from improved moderation without any corresponding enforcement mechanism. This asymmetry may distort platform incentives and lead to inefficiencies in content governance.

Finally, the role such infrastructure firms play blurs the boundary between market intermediaries and regulators. Distributors with the capacity to enforce content standards function simultaneously as infrastructure and as normative enforcers. This hybrid role challenges existing legal and policy frameworks, which often presume a clear division between infrastructure provision and content regulation.

In sum, the findings suggest that dominant distribution intermediaries can exert substantial influence over platform behaviour, not through direct involvement in content production or moderation, but through their gatekeeping position in the digital economy. This influence extends beyond structural incentives to shape the boundaries of political speech itself: platforms may suppress controversial but lawful content, such as election fraud claims or January 6 commentary, not because of regulatory obligation, but to avoid commercial risk. Regulatory frameworks that address not only platform-level decisions, but also the market structure and power of the intermediaries that govern access to users are therefore necessary to preserve a pluralistic marketplace of ideas.

6 Conclusion

This paper has shown that infrastructure providers, through their control of key distribution channels, can exert substantial influence over the governance of online speech. By conditioning platform access on compliance with content moderation standards, dominant app distributors like Google shape not only which platforms can reach users, but also the behaviour of users and the content that circulates within these ecosystems.

Using the July 2022 update to Google's Play Store policies as a quasi-experimental setting, I find that threatening content and politically sensitive misinformation declined sharply on platforms exposed to the new enforcement regime, while no comparable change occurred on an unexposed platform. These effects are concentrated among users who previously exhibited high levels of noncompliant content, and they persist over time. The results are robust to a range of alternative specifications and falsification exercises, strengthening the causal interpretation.

The findings have important implications for the political economy of digital markets. First, they demonstrate that private enforcement mechanisms can meaningfully shape the boundaries of online discourse even absent state intervention. Infrastructure providers, by leveraging control over market access, effectively set behavioural expectations for platforms and users, operating as de facto regulators without the procedural safeguards that typically constrain public authority.

Second, the governance-by-infrastructure model raises concerns about transparency, accountability, and legitimacy. Enforcement decisions are often opaque, driven by internal policies that are neither democratically ratified nor consistently applied. While infrastructure pressure can reduce the prevalence of harmful or dangerous speech, it also creates risks of overcorrection and inequitable treatment across platforms and communities.

Third, this form of regulation is mediated by market structure: platforms more dependent on dominant distributors have stronger incentives to comply, while those less reliant on mainstream infrastructure may resist. This uneven exposure complicates both enforcement outcomes and policy responses, suggesting that interventions focused solely on platform behaviour may miss a crucial upstream source of governance.

Finally, the findings invite a broader reconsideration of how we conceptualise governance in digital markets. Rather than treating platforms as isolated sites of decision-making, we must recognise them as embedded in complex supply chains where infrastructure providers, often operating behind the scenes, play a central role in shaping information flows.

Future research should further investigate the conditions under which private enforcement complements or substitutes for public regulation, the distributional consequences of infrastructure-led

governance, and the trade-offs between effectiveness, legitimacy, and innovation in moderating online speech. As control over digital pathways continues to consolidate, understanding how market power and normative power interact will be essential for designing policy frameworks that preserve the openness, accountability, and pluralism of the online public sphere.

References

Acemoglu, D. and Robinson, J. A. (2008). The role of institutions in growth and development. Technical Report Working Paper No. 10, Commission on Growth and Development.

Agarwal, S., Ananthakrishnan, U. M., and Tucker, C. E. (2022). Content moderation at the infrastructure layer: Evidence from parler. *Available at SSRN 4232871*.

Aliapoulios, M., Bevensee, E., Blackburn, J., De Cristofaro, E., and Zannettou, S. (2021). An early look at the parler online social network. arXiv preprint arXiv:2101.03820. `https://arxiv.org/abs/2101.03820`.

Allcott, H., Braghieri, L., Eichmeyer, S., and Gentzkow, M. (2020). The welfare effects of social media. *American Economic Review*, 110(3):629–676.

Allcott, H. and Gentzkow, M. (2017). Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31(2):211–236.

Aluru, S. S., Mathew, B., Saha, P., and Mukherjee, A. (2020). Deep learning models for multilingual hate speech detection. *arXiv preprint arXiv:2004.06465*.

Andres, R. and Slivko, O. (2021). Combating online hate speech: The impact of legislation on twitter. Technical report, ZEW Discussion Papers.

Aridor, G., Jiménez-Durán, R., Levy, R., and Song, L. (2024). The economics of social media. *Journal of Economic Literature*, 62(4):1422–1474.

Armstrong, M. (2006). Competition in two-sided markets. *RAND Journal of Economics*, 37(3):668–691.

Beknazar-Yuzbashev, G., Jiménez Durán, R., and Stalinski, M. (2024). A model of harmful yet engaging content on social media. *Available at SSRN*.

Berlinski, N., Doyle, M., Guess, A. M., Levy, G., Lyons, B., Montgomery, J. M., Nyhan, B., and Reifler, J. (2023). The effects of unsubstantiated claims of voter fraud on confidence in elections. *Journal of Experimental Political Science*, 10(1):34–49.

Besley, T. and Prat, A. (2006). Handcuffs for the grabbing hand? media capture and government accountability. *American Economic Review*, 96(3):720–736.

Cao, B., Lindo, J. M., and Zhong, J. (2023). Can social media rhetoric incite hate incidents? evidence from trump's "chinese virus" tweets. *Journal of Urban Economics*, 137:103580.

Dixon, L., Li, J., Sorensen, J., Thain, N., and Vasserman, L. (2018). Measuring and mitigating unintended bias in text classification. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 67–73.

Duran, V. J., Muller, K., and Schwarz, C. (2022). The effect of content moderation on online and offline hate: Evidence from germany's netzdg. Working Paper. Forthcoming .

Goodwin, C. N. and Woolley, S. (2022). Sideloading: An exploration of drivers and motivations. In *35th International BCS Human-Computer Interaction Conference*, pages 1–6. BCS Learning & Development.

Guess, A., Nyhan, B., and Reifler, J. (2020). Exposure to untrustworthy websites in the 2016 us election. *Nature Human Behaviour*, 4(5):472–480.

Hart, O. and Tirole, J. (1990). Vertical integration and market foreclosure. *Brookings Papers on Economic Activity: Microeconomics*, 1990:205–285.

Horta Ribeiro, M., Hosseinmardi, H., West, R., and Watts, D. J. (2023). Deplatforming did not decrease parler users' activity on fringe social media. *PNAS nexus*, 2(3):pgad035.

Jones, J. J., Bond, R. M., Bakshy, E., Eckles, D., and Fowler, J. H. (2017). Social influence and political mobilization: Further evidence from a randomized experiment in the 2012 us presidential election. *PloS one*, 12(4):e0173851.

Kalra, A. (2025). Hate in the time of algorithms: Evidence on online behavior from a large-scale experiment. *arXiv preprint arXiv:2503.06244*.

Klein, B. and Murphy, K. M. (1988). Vertical restraints as contract enforcement mechanisms. *Journal of Law and Economics*, 31(2):265–297.

Klonick, K. (2018). The new governors: The people, rules, and processes governing online speech. *Harvard Law Review*, 131(6):1598–1670.

Kominers, S. D. and Shapiro, J. M. (2024). Content moderation with opaque policies. Technical report, National Bureau of Economic Research.

Lafontaine, F. and Slade, M. (2007). Vertical integration and firm boundaries: The evidence. *Journal of Economic Literature*, 45(3):629–685.

Levy, R. (2021). Social media, news consumption, and polarization: Evidence from a field experiment. *American Economic Review*, 111(3):831–870.

Liu, Y., Yildirim, P., and Zhang, Z. J. (2021). Social media, content moderation, and technology. *arXiv preprint arXiv:2101.04618*.

Madio, L., Mitchell, M., Quinn, M., and Reggiani, C. (2025). Asymmetric content moderation in search markets: The case of adult websites. Technical report, CESifo Working Paper.

Madio, L. and Quinn, M. (2024). Content moderation and advertising in social media platforms. *Journal of Economics & Management Strategy*.

Mosquera, R., Galletta, S., and Tavits, M. (2020). Does social media reinforce mobilization? facebook, public events, and the 2016 us presidential election. *American Political Science Review*, 114(4):1325–1341.

Muller, K. and Schwarz, C. (2019). Fanning the flames of hate: Twitter and anti-minority violence. *American Economic Review*, 109(10):3594–3642.

Müller, K. and Schwarz, C. (2023). The effects of online content moderation: Evidence from president trump's account deletion. *Available at SSRN 4296306*.

Müller, K. and Schwarz, C. (2021). Fanning the flames of hate: Social media and hate crime. *Journal of the European Economic Association*, 19(4):2131–2167.

Prat, A. (2018). Media power. *Quarterly Journal of Economics*, 133(3):1331–1383.

Rey, P. and Tirole, J. (2001). Integration, delegation and the nature of contracts. *Quarterly Journal of Economics*, 116(4):1063–1100.

Rey, P. and Tirole, J. (2007). A primer on foreclosure. In Armstrong, M. and Porter, R., editors, *Handbook of Industrial Organization*, volume 3, pages 2145–2220. Elsevier.

Rey, P. and Vergé, T. (2004). Bilateral control with vertical contracts. *RAND Journal of Economics*, pages 728–746.

Rizzi, M. (2024). Self-regulation of social media and the evolution of content: a cross-platform analysis. *Available at SSRN 5018309*.

Roberts, S. T. (2019). *Behind the Screen: Content Moderation in the Shadows of Social Media*. Yale University Press, New Haven, CT.

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *Proceedings of the 26th international conference on world wide web*, pages 1391–1399.

Zannettou, S., Bradlyn, B., De Cristofaro, E., Sirivianos, M., Stringhini, G., and Blackburn, J. (2018). What is Gab? a bastion of free speech or an alt-right echo chamber? In *Companion Proceedings of the The Web Conference 2018*, pages 1007–1014. https://dl.acm.org/doi/10.1145/3184558.3191531.

Appendix

A Tables

Table A.1: Timeline of Google Play Policy Changes Relevant to Moderation Enforcement

Date	Policy	Key Provisions
Pre-2022	UGC Policy Baseline	Required moderation, terms of service compliance, in-app reporting and blocking systems.
Pre-2022	Hate Speech Baseline	Prohibited incitement and hate-based content; provided clarified examples.
Apr 6, 2022	Hate Speech Reiteration	Reaffirmed scope of incitement bans and compliance with local legal regimes, effective May 11 [Archive]
Apr 6, 2022	Expanded UGC Rules	Introduced system-level UGC requirements, effective October 11. [Archive]
Jul 27, 2022	Sensitive Events Clause	Enforceable immediately; required platforms to avoid insensitivity toward tragic events. [Archive]
Aug 19, 2022	Truth Social Rejection	Google formally rejected the app for inadequate content moderation. [News]
Oct 11, 2022	UGC Rule Enforcement	Formal deadline for compliance with the April UGC requirements.
Oct 12, 2022	Truth Social Acceptance	Google publicly announces acceptance of the app citing adequate content moderation. [News]

Table A.2: Comparison of Google's Sensitive Events Policy Before and After July 27, 2022

Policy Component	**Before (Jan 17, 2022)**	**After (July 27, 2022)**
Opening Clause	"lack reasonable sensitivity towards or capitalize on a natural disaster, atrocity, health crisis, conflict, death, or other tragic event"	"capitalize on or are insensitive toward a sensitive event with significant social, cultural, or political impact"
Definition of Sensitive Events	Examples included: natural disaster, atrocity, health crisis, conflict, death, or tragic event	Expanded and clarified: civil emergencies, natural disasters, public health emergencies, conflicts, deaths, or other tragic events
EDSA Clause	Allowed if content had Educational, Documentary, Scientific, or Artistic (EDSA) value or aimed to raise awareness	Same: EDSA clause retained verbatim
Event Denial Clause	"Denying a major tragic event"	"Denying the occurrence of a well-documented, major tragic event"
Profiteering Clause	Profiting from a tragic event without benefit to victims	Profiting from a sensitive event without benefit to victims

Table A.3: Descriptive statistics

Panel A: User Post Data	Gettr	TruthSocial	Gab
Users in Sample	25952	13586	7194
Likes (per user post)	10.900	3095.498	11.136
	[131.070]	[1.7e+04]	[97.994]
Reposts (per user post)	3.767	1123.812	5.769
	[53.188]	[4988.941]	[44.359]
Replies (per user post)	1.013	219.964	1.474
	[13.435]	[2061.993]	[10.747]
N_{pre}	877794	2687183	4160391
N_{tot}	1757312	11785676	13156911
Panel B: User Panel Data			
% Users post on a given day	0.127	0.296	0.343
	[0.332]	[0.456]	[0.475]
% Users comment on a given day	0.061	0.027	0.343
	[0.239]	[0.161]	[0.475]
Daily Posts	0.281	1.716	2.306
	[1.236]	[6.760]	[9.176]
Daily Likes of a Users posts	1.895	5310.825	25.652
	[58.978]	[3.0e+04]	[284.404]
Daily Replies to a Users posts	0.176	377.383	3.395
	[5.894]	[2934.263]	[28.819]
Daily Reposts of a Users posts	0.655	1928.080	13.288
	[23.341]	[9708.583]	[126.677]
N_{pre}	5050095	1566267	1806241
N_{tot}	28297340	13698512	8238494
Panel C: Post Measures			
Threat score	0.034	0.026	0.045
	[0.082]	[0.065]	[0.098]

Continued on next page

	Gettr	TruthSocial	Gab
Max Threat score	0.444	0.535	0.637
	[0.169]	[0.137]	[0.128]
Toxicity	0.175	0.144	0.193
	[0.210]	[0.175]	[0.215]
Severe toxicity	0.020	0.013	0.028
	[0.066]	[0.046]	[0.084]
Identity attack	0.043	0.032	0.069
	[0.091]	[0.070]	[0.132]
Insult	0.122	0.096	0.125
	[0.194]	[0.163]	[0.193]
Profanity	0.085	0.066	0.100
	[0.160]	[0.122]	[0.172]
Election fraud	0.020	0.039	0.019
	[0.139]	[0.194]	[0.137]
Anti-vaccine	0.044	0.024	0.067
	[0.204]	[0.153]	[0.251]
QAnon	0.018	0.023	0.044
	[0.132]	[0.151]	[0.205]
January 6	0.004	0.003	0.004
	[0.065]	[0.059]	[0.065]
Climate denial	0.001	0.001	0.002
	[0.032]	[0.033]	[0.040]

Note: This table summarises user-level content characteristics and content across platforms from before July 27, 2022. Panel A captures user level posts conditional on posting. Panel B utilises the panel dataset, which is at the user-day level. Panel C are average pre-treatment measures from the continuous Perspectives API scores and binary measures of sensitive event flags.

Table A.4: Examples of User Posts Threat Scores

Date	**Score**	**Post Content**
30 Jun 2022	0.88	I would've killed them all
01 Jul 2022	0.85	I say kill this communist!
26 Jul 2022	0.80	I will go personally kill the invaders.
16 Jul 2022	0.75	THIS MOTHER FUCKER NEEDS TO DIE!
08 Jun 2022	0.73	Why hasn't someone killed this piece of sht yet? Nazi trash deserves to be sent straight to hell.
07 Apr 2022	0.73	He needs a bullet in his head
07 Jun 2022	0.70	Soros is PURE EVIL. He needs to DIE!
19 Feb 2022	0.70	It's about this time I would have yanked him off the horse and kicked him in the face repeatedly
02 Mar 2022	0.60	"Wish I could of jumped through the screen and beat those wankers too a pulp, vile bastards"
19 Feb 2022	0.50	They don't want the pesky media there filming when they incite violence and start shooting people.
06 Jul 2022	0.40	Give up your guns at your peril people.
21 Mar 2022	0.30	Bill is one of the dumbest fcking people on television. He should go back to sticking his thumb up his ass
20 Nov 2021	0.20	Sudden Deaths of healthy young athletes. Vaccines are killing off Americans

Note: This table provides examples of users text contents from posts across Gab, Truth Social and Gettr and their coinciding 'Threat Score' as designated by the Perspectives API.

Table A.5: Selection Testing

	Gettr			**Truth Social**			**Gettr**		**Truth Social**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Threat	0.5454***	0.5142***	0.7285***	0.6654***	0.6095***	1.3059***	-7.9373***	-4.6341**	8.9951***	8.5292***
	(0.0668)	(0.0664)	(0.1219)	(0.1868)	(0.1779)	(0.2384)	(2.3016)	(2.2478)	(2.7252)	(2.6825)
N	47,249	47,249	47,249	18,562	18,562	18,562	26,033	26,033	13,586	13,586
# Leavers	16501	16501	16501	4886	4886	4886	-	-	-	-
# Stayers	30748	30748	30748	13676	13676	13676	30748	30748	13676	13676
User Controls	No	Yes	Yes	No	Yes	Yes	No	Yes	No	Yes
Weights	No	No	Yes	No	No	Yes	No	No	No	No

Note: The table presents two sets of regressions examining user selection and attrition relative to the policy change. Columns (1)–(6) estimate the likelihood that a user posts after July 27, 2022, as a function of pre-treatment threat scores. Columns (7)–(10) estimate the number of days users remained active after the policy. Specifications progressively add controls for user posting behaviour and exposure. Columns (3) and (6) apply analytic weights. All models use one observation per user. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A.6: Selection Testing (Gab)

	Gab				
	(1)	(2)	(3)	(4)	(5)
Threat	2.4201***	1.7977***	2.9157***	25.1216***	19.6488***
	(0.2049)	(0.1837)	(0.1976)	(4.7898)	(4.7119)
N	14,860	14,860	14,860	7,533	7,533
# Leavers	7327	7327	7327	-	-
# Stayers	7533	7533	7533	7533	7533
User Controls	No	Yes	Yes	No	Yes
Weights	No	No	Yes	No	No

Note: The table presents two sets of regressions examining user selection and attrition relative to the policy change. Columns (1)–(3) estimate the likelihood that a user posts after July 27, 2022, as a function of pre-treatment threat scores. Columns (4)–(5) estimate the number of days users remained active after the policy. Specifications progressively add controls for user posting behaviour and exposure. Column (3) applies analytic weights. All models use one observation per user. Robust standard errors in parentheses. *** p<0.01, ** p<0.05, * p<0.10.

Table A.7: Robustness I - Google Apps

	Gettr				**Truth Social**			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{Threat}$	-0.2196***				-0.1064***			
× Post	(0.0045)				(0.0105)			
$\overline{Threat}$		-0.0175***				-0.0052***		
(Top Quartile) × Post		(0.0004)				(0.0005)		
$\overline{Threat}$			-0.0272***				-0.0105***	
(Top Decile) × Post			(0.0009)				(0.0014)	
$\overline{Threat}$				-0.0065***				-0.0044***
(Q2) × Post				(0.0004)				(0.0006)
$\overline{Threat}$				-0.0137***				-0.0067***
(Q3) × Post				(0.0004)				(0.0006)
$\overline{Threat}$				-0.0268***				-0.0110***
(Q4) × Post				(0.0005)				(0.0008)
N	2,968,879	2,968,879	2,968,879	2,968,879	11466215	11466215	11466215	11466215
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	25952	25952	25952	25952	13515	13515	13515	13515
DV Pre-Mean	0.0336	0.0336	0.0336	0.0336	0.0265	0.0265	0.0265	0.0265
Treatment SD	0.0819	0.0819	0.0819	0.0819	0.0657	0.0657	0.0657	0.0657

Note: The table presents the results of estimating alternative specifications supporting robustness of Equation 1 results. Column (1) dependent variable is the natural logarithm of the user content threat score while columns (2) through (4) are level. In column (2), I replace the main variable of interest with pre-treatment average threat score quartiles compared to the lowest quartiles, and then a dummy for the top quartile and decile respectively in columns (3) and (4). The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.8: Robustness II - Google Apps

	Truth Social				Gettr			
	(1)	(2)	(3)	(4)	(1)	(2)	(3)	(4)
$\overline{Threat}$× Post (01apr2022)	-0.0080***				-0.0262***			
	(0.0012)				(0.0007)			
$\overline{Threat}$× Post (01may2022)		-0.0078***				-0.0219***		
		(0.0013)				(0.0007)		
$\overline{Threat}$× Post (01jun2022)			-0.0062***				-0.0171***	
			(0.0013)				(0.0010)	
$\overline{Threat}$× Post (01jul2022)				-0.0042***				-0.0131***
				(0.0012)				(0.0010)
N	4,747,170	4,733,070	4,689,190	4,588,967	716,627	687,676	646,510	594,740
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	10,533	10,019	8,813	7,532	13,445	11,766	10,203	8,591
DV Pre-Mean	0.0278	0.0278	0.0278	0.0278	0.0334	0.0334	0.0334	0.0334
Treatment SD	0.0633	0.0633	0.0633	0.0633	0.0805	0.0805	0.0805	0.0805

Note: The table presents the results of estimating Equation 1 with different cut off dates for calculating the pre-treatment average user threat score. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.9: Robustness III - Google Apps

	Gettr					Truth Social				
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Max Threat	-0.0082***					-0.0026***				
× Post	(0.0002)					(0.0003)				
Max Threat		-0.0150***					-0.0077***			
(≥ 0.2)× Post		(0.0003)					(0.0003)			
Max Threat			-0.0118***					-0.0043***		
(≥ 0.4)× Post			(0.0003)					(0.0003)		
Max Threat				-0.0103***					-0.0015***	
(≥ 0.6)× Post				(0.0008)					(0.0005)	
Max Threat					0.0078					-0.0024
(≥ 0.8)× Post					(0.0126)					(0.0016)
N	2,968,879	2,968,879	2,968,879	2,968,879	2,968,879	11466215	11466215	11466215	11466215	11466215
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	25952	25952	25952	25952	25952	13515	13515	13515	13515	13515
DV Pre-Mean	0.0336	0.0336	0.0336	0.0336	0.0336	0.0265	0.0265	0.0265	0.0265	0.0265
Treatment SD	0.1698	0.1698	0.1698	0.1698	0.1698	0.1366	0.1366	0.1366	0.1366	0.1366

Note: The table presents the results of estimating Equation 1 with alternative definitions for how threatening users were in the pre-treatment period based on their maximum threat score, rather than on their average pre-treatment threat score. Column (1) uses a continuous variable while columns (2) though (4) are dummy variables equal to 1 at and above the stated threshold and 0 otherwise. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.10: Robustness IV - Google Apps

	Gettr			**Truth Social**		
	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0144***	-0.0137***	-0.0167***	-0.0072***	-0.0069***	-0.0029***
× Post	(0.0008)	(0.0006)	(0.0008)	(0.0007)	(0.0013)	(0.0010)
Source	Original	Reply	Repost	Original	Reply	Repost
N	710,427	1,217,429	1,035,971	7,964,599	256,713	3,249,490
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	20572	20971	19163	13347	1590	6965
DV Pre-Mean	0.0322	0.0346	0.0330	0.0263	0.0288	0.0262
Treatment SD	0.0778	0.0880	0.0751	0.0651	0.0753	0.0622

Note: The table presents the results of estimating Equation 1 split by the types of content. Columns (1) are original posts content (not comments or replies); columns (2) are replies/comments; and columns (3) are re-posts of other users material. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.11: Robustness V - Google Apps

	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0181***	-0.0936***	-0.0066***	-0.0121***	-0.0569***	-0.0355***
× Post	(0.0005)	(0.0034)	(0.0004)	(0.0010)	(0.0027)	(0.0020)
N	2,968,879	2,968,879	2,968,879	2,968,879	2,968,879	2,968,879
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Control × Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	25952	25952	25952	25952	25952	25952
DV Pre-Mean	0.0336	0.1749	0.0200	0.0431	0.1216	0.0848
Treatment SD	0.0819	0.2103	0.0658	0.0911	0.1943	0.1598
	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0082***	-0.0052***	-0.0004	-0.0038***	-0.0048***	-0.0006
× Post	(0.0006)	(0.0018)	(0.0013)	(0.0008)	(0.0016)	(0.0016)
N	11466215	11466215	11466215	11466215	11466215	11466215
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Control × Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	13515	13515	13515	13515	13515	13515
DV Pre-Mean	0.0265	0.1465	0.0133	0.0324	0.0978	0.0670
Treatment SD	0.0657	0.1772	0.0471	0.0711	0.1648	0.1236

Note: The table presents the results of estimating Equation 1 split by the types of content. Columns (1) are original posts content (not comments or replies); columns (2) are replies/comments; and columns (3) are re-posts of other users material. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.12: Robustness VI - Google Apps

	Gettr		**Truth Social**		**Gettr**		**Truth Social**	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
$\overline{Threat}$	-0.0202***	-0.0226***	-0.0103***	-0.0101***	-0.0269***	-0.0269***	-0.0211***	-0.0211***
× Post	(-37.96)	(-17.02)	(-13.19)	(-9.06)	(-60.67)	(-60.61)	(-14.76)	(-14.76)
N	3,490,683	754,169	11,914,040	8,488,525	3,493,325	3,136,067	11,915,521	11,785,676
# of Users	36305	5091	16891	4000	38944	26033	18371	13586
User FE	Yes	Yes	Yes	Yes	No	No	No	No
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	No	No	No	No
Endline Sample	No	Yes	No	Yes	-	-	-	-
Panel Sample	-	-	-	-	No	Yes	No	Yes

Note: This table reports robustness checks addressing potential bias from selective attrition and panel sample selection. Columns (1)–(4) restrict the sample to users observed through the end of the study period ("endline" sample), testing for bias due to differential post-treatment attrition. Columns (5)–(8) estimate a simplified specification without user fixed effects to include users not present throughout, comparing the full sample to the subset of "stayers" who are active both before and after the policy change. All regressions include year-day fixed effects. Standard errors clustered at the user level. *** p<0.01, ** p<0.05, * p<0.10.

Table A.13: Robustness I - Gab

	(1)	(2)	(3)	(4)
$\overline{Threat}$	-0.0399***			
× Post	(0.0120)			
$\overline{Threat}$		-0.0049***		
(Top Quartile) × Post		(0.0011)		
$\overline{Threat}$			-0.0066***	
(Top Decile) × Post			(0.0020)	
$\overline{Threat}$				-0.0014
(Q2) × Post				(0.0013)
$\overline{Threat}$				-0.0018
(Q3) × Post				(0.0012)
$\overline{Threat}$				-0.0064***
(Q4) × Post				(0.0015)
N	13501666	13501666	13501666	13501666
User FE	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
# of Users	7384	7384	7384	7384
DV Pre-Mean	0.0441	0.0441	0.0441	0.0441
Treatment SD	0.0980	0.0980	0.0980	0.0980

Note: The table presents the results of estimating alternative specifications supporting robustness of Equation 1 results. Column (1) dependent variable is the natural logarithm of the user content threat score while columns (2) through (4) are level. In column (2), I replace the main variable of interest with pre-treatment average threat score quartiles compared to the lowest quartiles, and then a dummy for the top quartile and decile respectively in columns (3) and (4). The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.14: Robustness II - Gab

	(1)	(2)	(3)	(4)
$\overline{Threat}$	-0.0022**			
(01apr2022) × Post	(0.0011)			
$\overline{Threat}$		-0.0019*		
(01may2022) × Post		(0.0011)		
$\overline{Threat}$			-0.0015	
(01jun2022) × Post			(0.0012)	
$\overline{Threat}$				-0.0016
(01jul2022) × Post				(0.0012)
N	13369475	13215525	13104253	12910241
User FE	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes
# of Users	6410	6129	5693	4968
DV Pre-Mean	0.0441	0.0441	0.0441	0.0441
Treatment SD	0.0980	0.0980	0.0980	0.0980

Note: The table presents the results of estimating Equation 1 with different cut off dates for calculating the pre-treatment average user threat score. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.15: Robustness III - Gab

	(1)	(2)	(3)	(4)	(5)
Max Threat	-0.0008				
× Post	(0.0007)				
Max Threat		-0.0046***			
(≥ 0.2)× Post		(0.0006)			
Max Threat			-0.0020**		
(≥ 0.4)× Post			(0.0010)		
Max Threat				-0.0001	
(≥ 0.6)× Post				(0.0009)	
Max Threat					-0.0010
(≥ 0.8)× Post					(0.0022)
N	13501666	13501666	13501666	13501666	13501666
User FE	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes
Controls	Yes	Yes	Yes	Yes	Yes
# of Users	7384	7384	7384	7384	7384
DV Pre-Mean	0.0441	0.0441	0.0441	0.0441	0.0441
Treatment SD	0.1230	0.1230	0.1230	0.1230	0.1230

Note: The table presents the results of estimating Equation 1 with alternative definitions for how threatening users were in the pre-treatment period based on their maximum threat score, rather than on their average pre-treatment threat score. Column (1) uses a continuous variable while columns (2) though (4) are dummy variables equal to 1 at and above the stated threshold and 0 otherwise. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.16: Robustness IV - Gab

	(1)	(2)
$\overline{Threat}$	0.0000	-0.0035***
× Post	(.)	(0.0011)
Source	Original	Reply
N	226	13501191
User FE	Yes	Yes
Year-Day FE	Yes	Yes
Controls	Yes	Yes
# of Users	11	7375
DV Pre-Mean	0.0405	0.0441
Treatment SD	0.0668	0.0980

Note: The table presents the results of estimating Equation 1 split by the types of content. Columns (1) are original posts content (not comments or replies); columns (2) are replies/comments; and columns (3) are re-posts of other users material. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.17: Robustness V - Gab

	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0035***	-0.0053***	-0.0023	0.0001	-0.0043***	-0.0035**
× Post	(0.0011)	(0.0019)	(0.0017)	(0.0017)	(0.0014)	(0.0017)
N	13204938	13204938	13204938	13204938	13204938	13204938
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Control × Time FE	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	7374	7374	7374	7374	7374	7374
DV Pre-Mean	0.0448	0.1928	0.0285	0.0689	0.1255	0.1005
Treatment SD	0.0979	0.2152	0.0842	0.1326	0.1932	0.1724

Note: The table presents the results of estimating Equation 1 split by the types of content. Columns (1) are original posts content (not comments or replies); columns (2) are replies/comments; and columns (3) are re-posts of other users material. The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.18: Robustness VI - Gab

	Gab			
	(1)	(2)	(3)	(4)
$\overline{Threat}$	-0.00336**	-0.00318***	-0.00564***	-0.00569***
× Post	(-2.52)	(-3.77)	(-4.88)	(-4.92)
N	16128850	11899570	16130864	15659334
# of Users	12846	3187	14860	7533
User FE	Yes	Yes	No	No
Year-Day FE	Yes	Yes	Yes	Yes
Controls	Yes	Yes	No	No
Endline Sample	No	Yes	-	-
Panel Sample	-	-	No	Yes

Note: This table reports robustness checks addressing potential bias from selective attrition and panel sample selection. Columns (1)–(2) restrict the sample to users observed through the end of the study period ("endline" sample), testing for bias due to differential post-treatment attrition. Columns (3)–(4) estimate a simplified specification without user fixed effects to include users not present throughout, comparing the full sample to the subset of "stayers" who are active both before and after the policy change. All regressions include year-day fixed effects. Standard errors clustered at the user level. *** p<0.01, ** p<0.05, * p<0.10.

Table A.19: Robustness I - DDD

	Platforms				Google			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gettr	-0.1814***							
$\overline{Threat}$ × Post	(0.0089)							
TS	-0.0738***							
$\overline{Threat}$ × Post	(0.0120)							
Gettr		-0.3860***						
$\overline{Threat}$ (Top Quartile) × Post		(0.0246)						
TS		-0.0980***						
$\overline{Threat}$ (Top Quartile) × Post		(0.0287)						
Gettr			-0.6045***					
$\overline{Threat}$ (Top Decile) × Post			(0.0435)					
TS			-0.2769***					
$\overline{Threat}$ (Top Decile) × Post			(0.0637)					
Gettr				-0.1690***				
$\overline{Threat}$ (Q2) × Post				(0.0278)				
TS				-0.1477***				
$\overline{Threat}$ (Q2) × Post				(0.0346)				
Gettr				-0.3383***				
$\overline{Threat}$ (Q3) × Post				(0.0251)				
TS				-0.2030***				
$\overline{Threat}$ (Q3) × Post				(0.0321)				
Gettr				-0.6193***				
$\overline{Threat}$ (Q4) × Post				(0.0314)				
TS				-0.2819***				
$\overline{Threat}$ (Q4) × Post				(0.0398)				
Google					-0.1050***			
$\overline{Threat}$ × Post					(0.0097)			
Google						-0.1744***		
$\overline{Threat}$ (Top Quartile) × Post						(0.0229)		
Google							-0.3689***	
$\overline{Threat}$ (Top Decile) × Post							(0.0465)	
Google								-0.1820***
$\overline{Threat}$ (Q2) × Post								(0.0232)
Google								-0.2523***
$\overline{Threat}$ (Q3) × Post								(0.0213)
Google								-0.3953***
$\overline{Threat}$ (Q4) × Post								(0.0284)
N	25217223	25217118	25217118	25217118	15540919	15540824	15540824	15540824
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform × Year-Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	46109	46106	46106	46106	45142	45139	45139	45139
DV Pre-Mean	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349
Treatment SD	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826

Note: The table presents the results of estimating alternative specifications supporting robustness of Equation 1 results. Column (1) dependent variable is the natural logarithm of the user content threat score while columns (2) through (4) are level. In column (2), I replace the main variable of interest with pre-treatment average threat score quartiles compared to the lowest quartiles, and then a dummy for the top quartile and decile respectively in columns (3) and (4). The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user.

Table A.20: Robustness II - DDD

	Platforms				Google			
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Gettr	-0.4705***							
$\overline{Threat}$ (April 2022) × Post	(0.0276)							
TS	-0.2155***							
$\overline{Threat}$ (April 2022) × Post	(0.0342)							
Gettr		-0.3484***						
$\overline{Threat}$ (May 2022) × Post		(0.0305)						
TS		-0.2114***						
$\overline{Threat}$ (May 2022) × Post		(0.0362)						
Gettr			-0.2208***					
$\overline{Threat}$ (June 2022) × Post			(0.0254)					
TS			-0.1420***					
$\overline{Threat}$ (June 2022) × Post			(0.0354)					
Gettr				-0.1107***				
$\overline{Threat}$ (July 2022) × Post				(0.0225)				
TS				-0.0327				
$\overline{Threat}$ (July 2022) × Post				(0.0288)				
Google					-0.3348***			
$\overline{Threat}$ (April 2022) × Post					(0.0270)			
Google						-0.3017***		
$\overline{Threat}$ (May 2022) × Post						(0.0275)		
Google							-0.1896***	
$\overline{Threat}$ (June 2022) × Post							(0.0267)	
Google								-0.0641***
$\overline{Threat}$ (July 2022) × Post								(0.0231)
N	24893092	24639339	24252769	23366504	15394356	15282411	15078100	14582049
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform × Year-Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	41740	39269	35149	28503	41535	39126	35051	28461
DV Pre-Mean	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349
Treatment SD	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826	0.0826

Note: The table presents the results of estimating alternative specifications supporting robustness of Equation 1 results. Column (1) dependent variable is the natural logarithm of the user content threat score while columns (2) through (4) are level. In column (2), I replace the main variable of interest with pre-treatment average threat score quartiles compared to the lowest quartiles, and then a dummy for the top quartile and decile respectively in columns (3) and (4). The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.21: Robustness III - DDD

	Platforms			Google						
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)
Gettr Max Threat × Post	-0.2026***									
	(0.0153)									
TS Max Threat × Post	-0.0702***									
	(0.0177)									
Gettr Max Threat (≥ 0.2) × Post		-0.3218***								
		(0.0166)								
TS Max Threat (≥ 0.2) × Post		-0.1849***								
		(0.0179)								
Gettr Max Threat (≥ 0.4) × Post			-0.2912***							
			(0.0290)							
TS Max Threat (≥ 0.4) × Post			-0.1274***							
			(0.0290)							
Gettr Max Threat (≥ 0.6) × Post				-0.2793***						
				(0.0312)						
TS Max Threat (≥ 0.6) × Post				-0.0410						
				(0.0268)						
Gettr Max Threat (≥ 0.8) × Post					0.1092					
					(0.3582)					
TS Max Threat (≥ 0.8) × Post					-0.0506					
					(0.0778)					
Google Max Threat × Post						-0.1105***				
						(0.0141)				
Google Max Threat (≥ 0.2) × Post							-0.2344***			
							(0.0145)			
Google Max Threat (≥ 0.4) × Post								-0.1710***		
								(0.0285)		
Google Max Threat (≥ 0.6) × Post									-0.0758***	
									(0.0210)	
Google Max Threat (≥ 0.8) × Post										-0.0450
										(0.0679)
N	25217118	25217118	25217118	25217118	25217118	15540824	15540824	15540824	15540824	15540824
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Platform × Year-Week FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	46106	46106	46106	46106	46106	45139	45139	45139	45139	45139
DV Pre-Mean	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349	0.0349
Treatment SD	0.1537	0.1537	0.1537	0.1537	0.1537	0.1537	0.1537	0.1537	0.1537	0.1537

Note: The table presents the results of estimating alternative specifications supporting robustness of Equation 1 results. Column (1) dependent variable is the natural logarithm of the user content threat score while columns (2) through (4) are level. In column (2), I replace the main variable of interest with pre-treatment average threat score quartiles compared to the lowest quartiles, and then a dummy for the top quartile and decile respectively in columns (3) and (4). The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

Table A.22: Robustness V - DDD

	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0473***	-0.0229***	-0.0350	-0.0128	-0.0325***	-0.0272***
× Post	(0.0150)	(0.0069)	(0.0256)	(0.0165)	(0.0085)	(0.0085)
Gettr	-0.4386***	-0.4533***	-0.2812***	-0.2511***	-0.3862***	-0.3577***
$\overline{Threat}$ × Post	(0.0227)	(0.0226)	(0.0372)	(0.0343)	(0.0261)	(0.0292)
TS	-0.2533***	0.0031	0.0117	-0.0752**	0.0062	0.0443*
$\overline{Threat}$ × Post	(0.0292)	(0.0132)	(0.0956)	(0.0301)	(0.0178)	(0.0242)
N	18267806	18267806	18267806	18267806	18267806	18267806
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	Yes	Yes	Yes	Yes	Yes	Yes
Platform × Year-Week FE	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	45701	45701	45701	45701	45701	45701
DV Pre-Mean	0.0349	0.1694	0.0202	0.0493	0.1126	0.0827
Treatment SD	0.0826	0.1989	0.0668	0.1052	0.1818	0.1496

	(1)	(2)	(3)	(4)	(5)	(6)
$\overline{Threat}$	-0.0473***	-0.0229***	-0.0350	-0.0128	-0.0325***	-0.0272***
× Post	(0.0150)	(0.0069)	(0.0256)	(0.0165)	(0.0085)	(0.0085)
Google	-0.3155***	-0.0095	-0.0695	-0.1082***	-0.0076	0.0207
$\overline{Threat}$ × Post	(0.0230)	(0.0132)	(0.0684)	(0.0270)	(0.0177)	(0.0232)
N	18267806	18267806	18267806	18267806	18267806	18267806
User FE	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes
Platform FE	Yes	Yes	Yes	Yes	Yes	Yes
User × Platform FE	Yes	Yes	Yes	Yes	Yes	Yes
Platform × Year-Week FE	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	45701	45701	45701	45701	45701	45701
DV Pre-Mean	0.0349	0.1694	0.0202	0.0493	0.1126	0.0827
Treatment SD	0.0826	0.1989	0.0668	0.1052	0.1818	0.1496

Note: The level of observation is user × year-day. All regressions control for user and year-day fixed effects. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

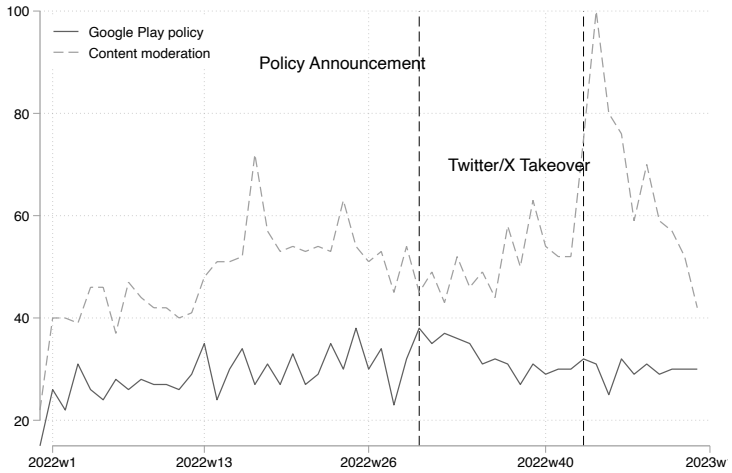
Table A.23: Spillovers

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
$\overline{Threat}$	-0.0039***	-0.0054***	-0.0039***				
× Post	(0.0014)	(0.0016)	(0.0014)				
Cross-Platform			-0.0022***				
× Post			(0.0007)				
Cross-Platform			-0.0015				
× $\overline{Threat}$ × Post			(0.0021)				
Gettr				-0.0005	0.0014		
$\overline{Threat}$ × Post				(0.0006)	(0.0023)		
Gettr Leaver					0.0009		
× Post					(0.0014)		
Gettr					-0.0025		
$\overline{Threat}$ × Gettr Leaver × Post					(0.0024)		
Cross-Platform						0.0061	-0.0153***
						(0.0074)	(0.0058)
Cross-Platform							0.0001
× Months Since Post							(0.0004)
N	11253525	2,004,577	13258102	950,085	950,085	539,085	504,137
User FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Year-Day FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Controls × Time FE	Yes	Yes	Yes	Yes	Yes	Yes	Yes
# of Users	4642	2555	7197	1516	1516	978	957
DV Pre-Mean	0.0456	0.0405	0.0448	0.0405	0.0405	.	.
Treatment SD	0.0990	0.0922	0.0979	0.0922	0.0922	.	.

Note: The table presents the results of estimating Equation 1. The dependent variable is the threat score of the users posts on Truth Social. $\overline{Threat}_i$ is the pre-treatment average threat score of user i's posts. User post threat score is standardised. $Post_t$ is an indicator variable = 1 following Google's announcement on August 19, 2022, and 0 otherwise. The level of observation is user × year-day. Standard errors clustered by user. *** p<0.01, ** p<0.05, * p<0.10.

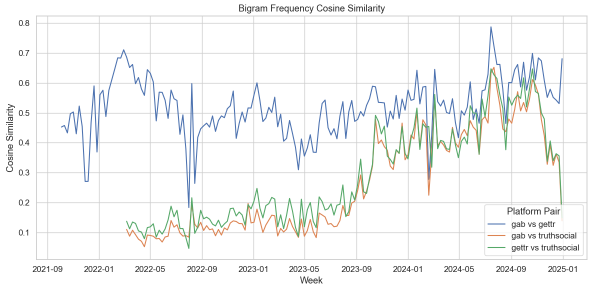
B Figures

Figure B.1: Google Trends - Google Policy and Content Moderation

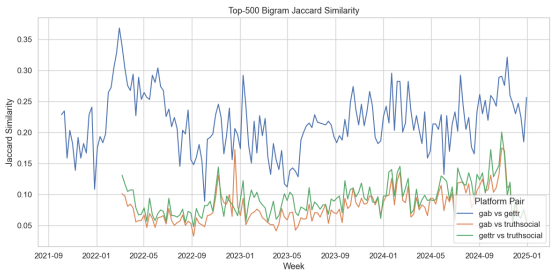


Note: Weekly Google Trends search interest for "Google Play policy" and "content moderation" worldwide. The Y-axis is a search pressure odds ration where 100 is the week with the ihghest serach of the associated term, while a scors of 50 means 50% of the search pressure compared to the highest search pressure.

Figure B.2: Similarity in content between platforms over time



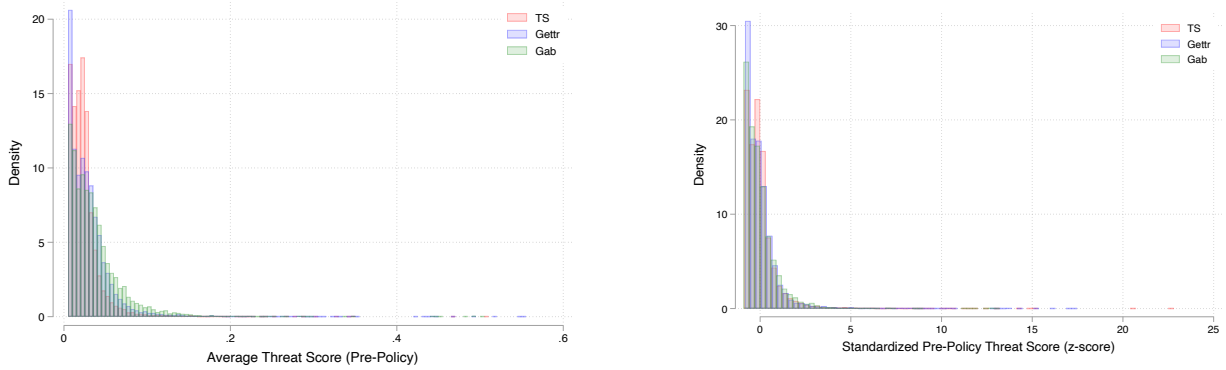
(a) Cosine Similarity



(b) Jaccard Similarity

Notes: This figure shows weekly pairwise similarity in language use between Gab, Gettr, and TruthSocial, based on n-gram frequency distributions. Panel (a) reports cosine similarity and panel (b) reports Jaccard similarity, computed from platform-specific n-gram distributions for each week. Higher values indicate greater overlap in the linguistic content of posts across platforms. The vertical line marks July 27, 2022, corresponding to Google's announced moderation policy change. Post-treatment convergence in similarity is visible, particularly between TruthSocial and the other platforms.

Figure B.3: Distribution of Pre-treatment Threat Scores Across Platforms

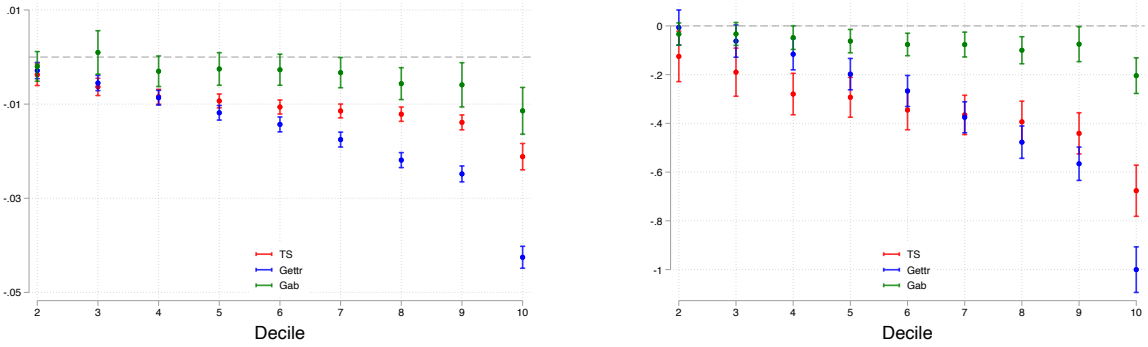


(a) Raw Distribution

(b) Z-Score Distribution

Notes: This figure compares the distribution of pre-treatment threat scores across platforms. Panel (a) shows the raw distribution of average threat scores by user, while panel (b) standardizes scores within each platform to highlight relative dispersion and tail behavior.

Figure B.4: Google Announcement on User Threatening Content Across Deciles Wiithin Platforms

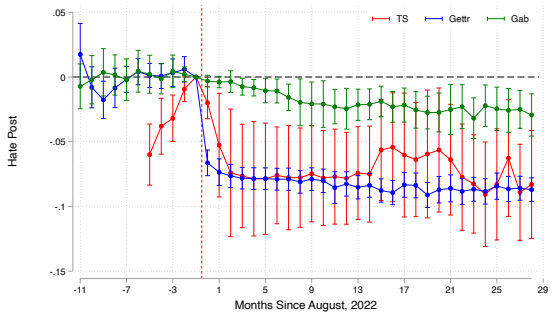


(a) Absolute Magnitude

(b) Mean Relative - Short Term Effect

Notes: Figure shows platform-specific regressions of the change in users' daily threat scores on a set of indicators for users' cross-sectional pre-treatment decile of average threat score. The omitted category is the first decile. In both panels, regressions include user control variables interacted with year-day fixed effects, user fixed effects, and year-day fixed effects. Standard errors are clustered at the user level. *** p<0.01, ** p<0.05, * p<0.10.

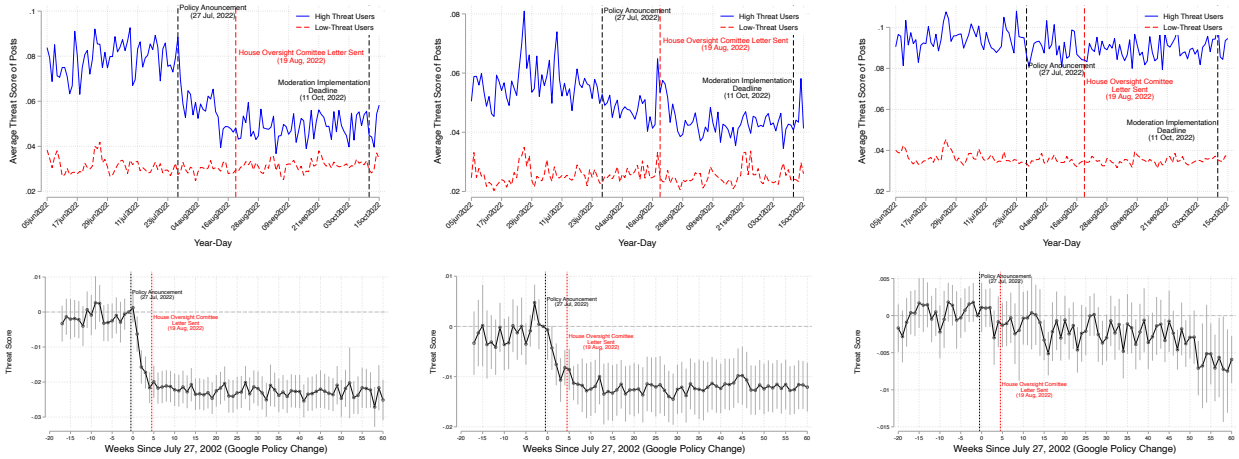
Figure B.5: Effect of Google Announcement on User Hate Speech Across Platforms



(a) Decile Regressions

Notes: Figure shows platform-specific regressions of users' posts classified as hate speech on the interaction of standardised pre-treatment user average hate speech prevalence with a set of month indicators relative to Google's announced policy change. The omitted category is July 2022. Standard errors are clustered at the user level. *** p<0.01, ** p<0.05, * p<0.10.

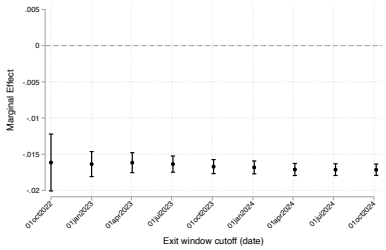
Figure B.6: User Threat Scores in days/weeks around July 27, 2022.



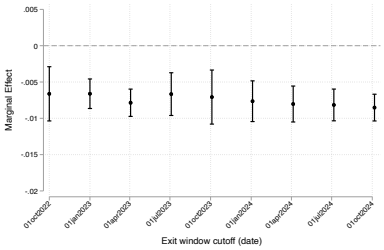
(a) Gettr **(b)** TS **(c)** Gab

Note: Across platform user threatening posts before and after Google announcement on July 27, 2022. The upper panels are the average threat score of posts by platform users, with users split based on the threat score of all of the users posts on that platform in the pre-exposure period. "High-Threat Users" are those whose posts before the Google announcement were on average above the 90th percentile of the threat score distribution. "Low-Threat Users" are all remaining users. The shaded areas indicate 95% confidence intervals for the means. Lower panels are platform specific regressions of user post threat score on the interaction of pre-treatment user average threat score cross-sections with a set of indicators for week pre-/ post- Google announcement. The omitted category is the week before the July 27, 2022. Standard errors are clustered at the user level. Regressions contain user control variables × year-day FE, user FE, year-day FE. Standard errors are clustered at the user level.

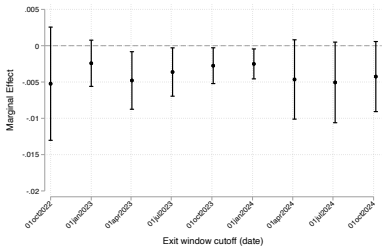
Figure B.7: Robustness - Selective Attrition



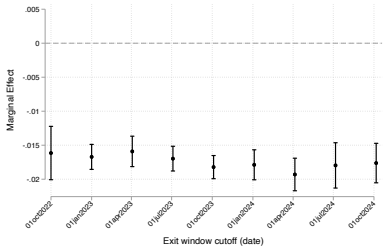
(a) Gettr - Accumulative



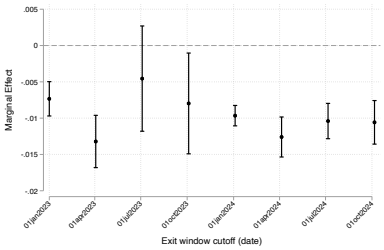
(b) TS - Accumulative



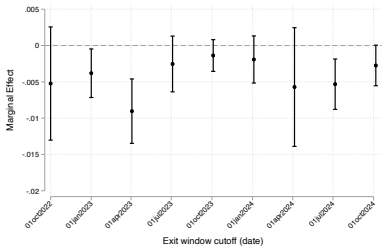
(c) Gab - Accumulative



(a) Gettr - Buckets



(b) TS - Buckets

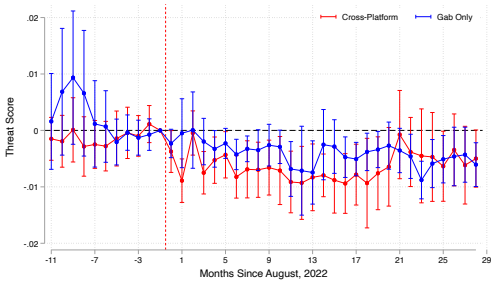


(c) Gab - Buckets

Note: This figure evaluates robustness to selective attrition by estimating the marginal effect of the post-policy interaction with pre-treatment threat level across exit cohorts. Each point represents the coefficient on the interaction term from a separate regression of user 'threateningness' on post × threat, restricted to users who exited the platform before a given quarterly cutoff date (top row), or who exited within non-overlapping 3-month buckets preceding each cutoff (bottom row). Vertical lines show 95% confidence intervals. All regressions are restricted to users observed both before and after treatment (i.e., "stayers"), and include user fixed effects, day fixed effects, and time-varying controls. Standard errors are clustered at the user level.

Figure B.8: Cross platform spillover event studies

(a) Gettr vs Gab



(b) Gettr users on Gab

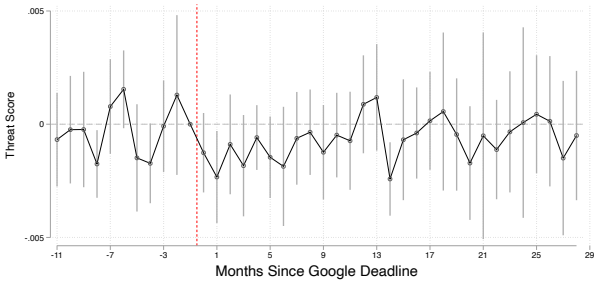
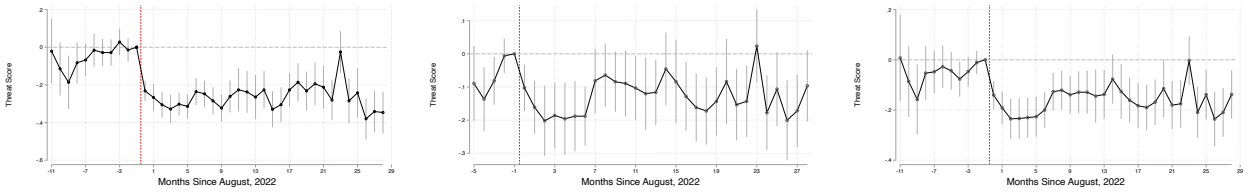


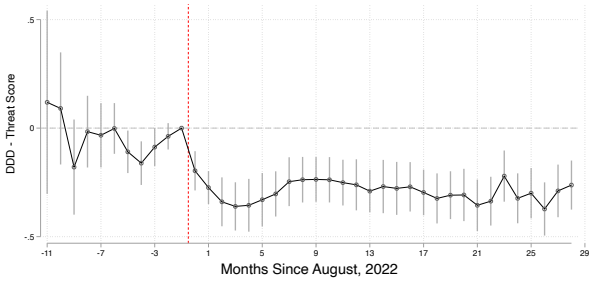
Figure B.9: DiD High Threat Users between platforms.



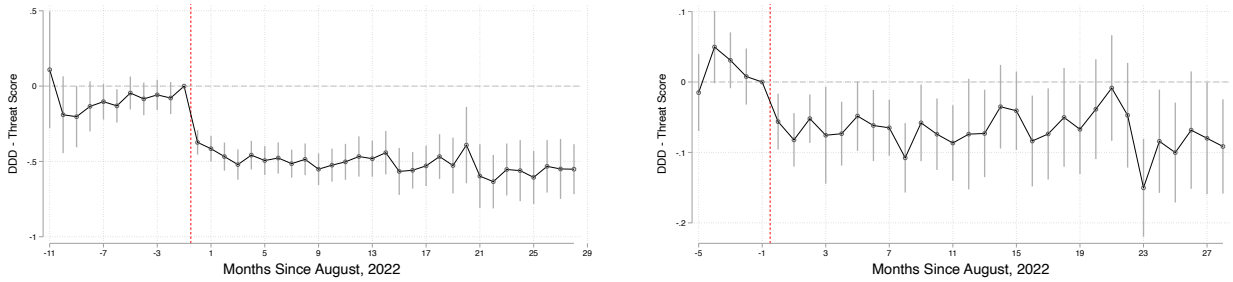
(a) Gettr **(b)** TS **(c)** Google

Note: Panels are regressions of user threatening score on the interaction of the Google exposed platform indicator with Gab as the omitted category, interacted with a set of indicators for month pre-/ post- Google announcement. The omitted category is July, 2022. Standard errors are clustered at the user level. Regressions contain user FE, platform-year-day FE and user-platform FE. Standard errors are clustered at the user level.

Figure B.10: DDD Estimates on Matched Sample by Pre-Treatment Threat Scores



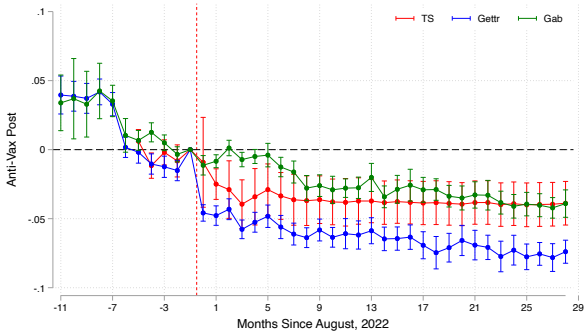
(a) Google



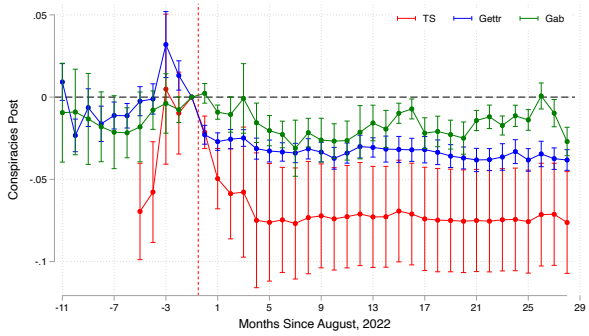
(b) Gettr **(c)** TS

Note: Each panel reports estimates from a difference-in-differences regression comparing changes in user threat scores between Gab and either the pooled Google-exposed platforms (Panel a), Gettr (Panel b), or (Panel c), using a sample matched on the distribution of pre-treatment threat scores. Users were matched by coarsened bins of standardized pre-treatment threat scores to ensure comparable threat distributions across platforms. Estimates show deviations from the pre-policy baseline (July 2022), with user and platform-day fixed effects included. Standard errors are clustered at the user level.

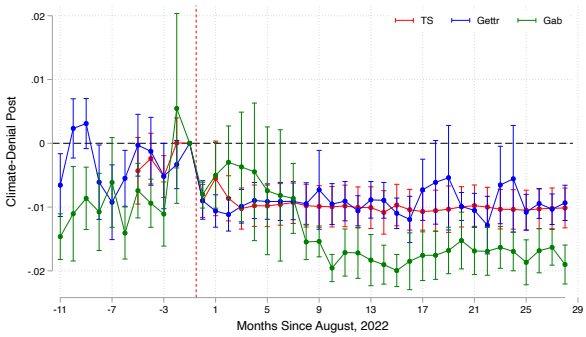
Figure B.11: Topic-specific event study estimates by pre-treatment topic intensity.



(a) Anti-vaccine content



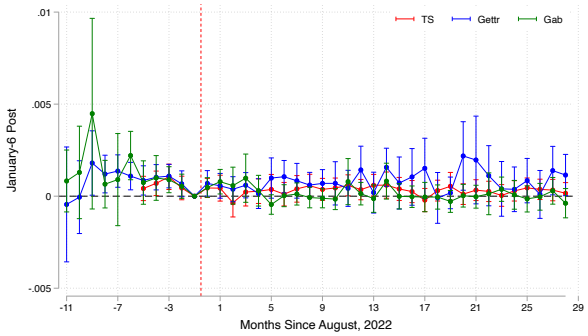
(b) Conspiracy theories



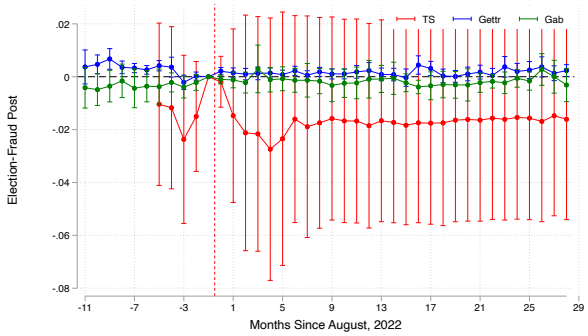
(c) Climate change denial

Note: Each panel shows the interaction effect between user-level pre-policy topic engagement (standardised) and time relative to Google's July 2022 policy update, using the dynamic difference-in-differences specification from Equation 1. Bars are 95% confidence intervals. Models include user fixed effects and platform-by-month trends. Standard errors clustered at the user level.

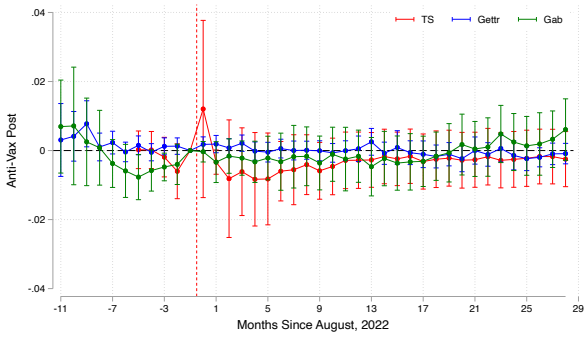
Figure B.12: Topic-specific event study falsification estimates by pre-treatment threatening average score.



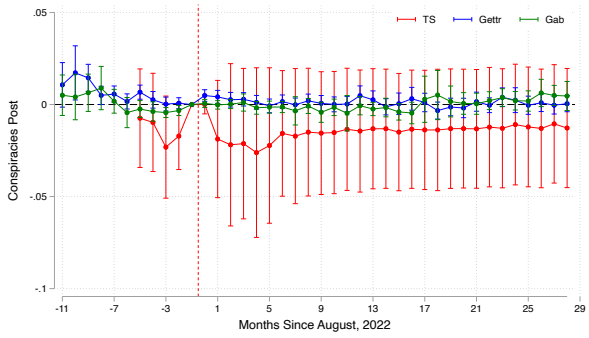
(a) January 6 commentary



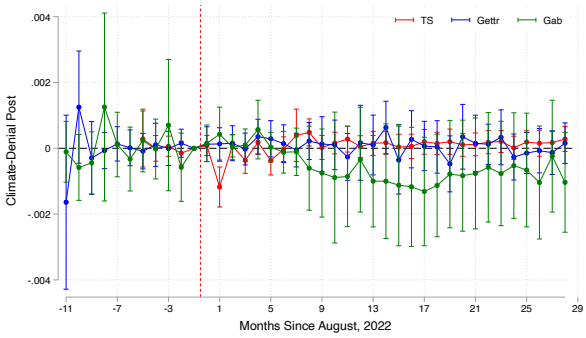
(b) Election fraud claims



(c) Anti-vaccine content



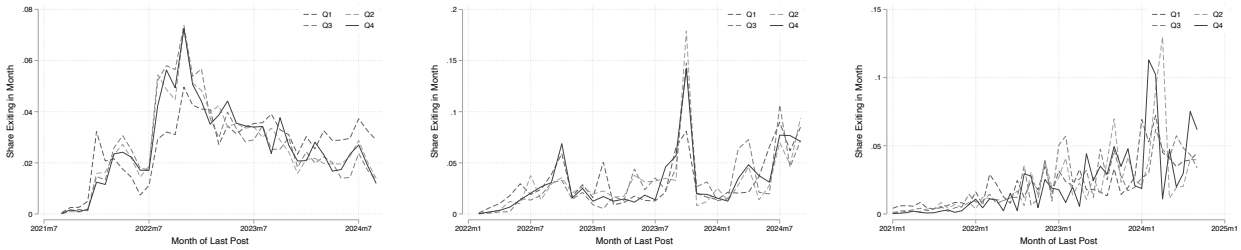
(d) Conspiracy theories



(e) Climate change denial

Note: Each panel shows the interaction effect between user-level pre-policy topic engagement (standardised) and time relative to Google's July 2022 policy update, using the dynamic difference-in-differences specification from Equation 1. Bars are 95% confidence intervals. Models include user fixed effects and platform-by-month trends. Standard errors clustered at the user level.

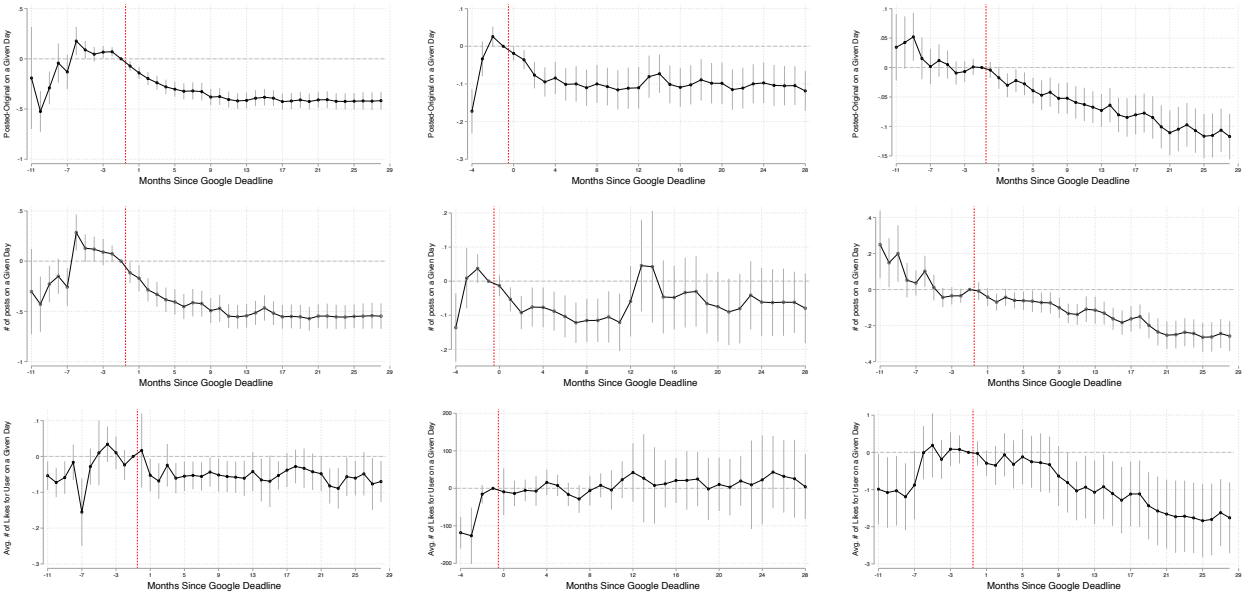
Figure B.13: Time Series of Platform Exiters by User 'Threateningess'



(a) Gettr **(b)** TS **(c)** Gab

Note: Each panel plots the monthly share of users who exited the platform (i.e., made their final post) by quartile of pre-treatment threateningness. Users are assigned to quartiles based on their average threatening score prior to the Google announcement. For each quartile, the share of users exiting in a given month is calculated as the number of users whose final post falls in that month divided by the total number of users in that quartile. The sample is restricted to users whose last post was before October 2024. Line types distinguish quartiles, with Q4 representing the highest baseline threat level.

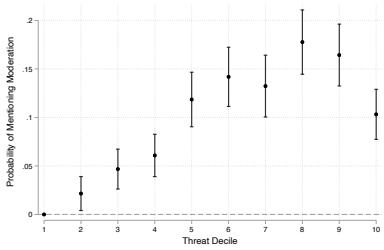
Figure B.14: User Engagement by Pre-Treatment Threat Intenstiy



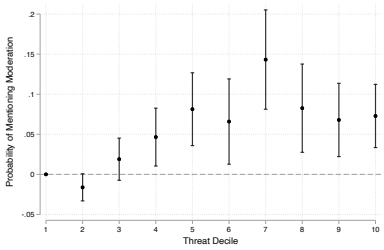
(a) Gettr **(b)** TS **(c)** Gab

Note: Panels are regressions of outcomes on user pre-treatment threatening score interacted with a set of indicators for month pre-/ post-Google announcement. The omitted category is July, 2022. The top panel outcome variable is an indicator for whether the user posted roginal content on a given day. The middle panel outcome is the number of posts a user makes on a given day (conditional on having joined the platform). The lower panel outcome variable is the number of like a post receives. Standard errors are clustered at the user level. Regressions contain user FE and year-day fixed effects. Standard errors are clustered at the user level.

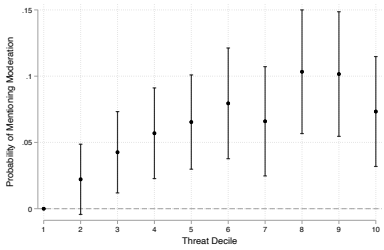
Figure B.15: Moderation Discourse Before User Disengagement Before and After Policy Change



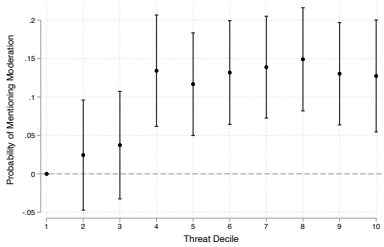
(a) Gettr - Before



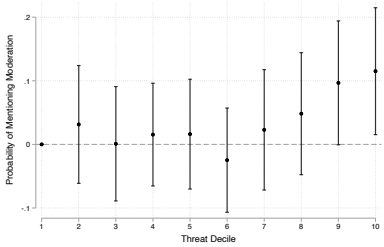
(b) TS - Before



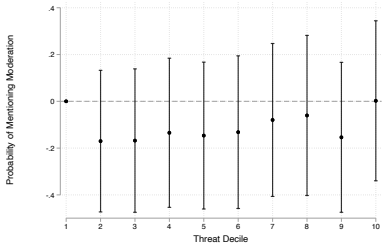
(c) Gab - Before



(a) Gettr - After



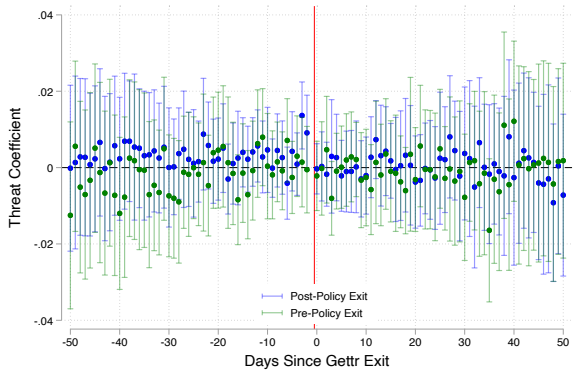
(b) TS - After



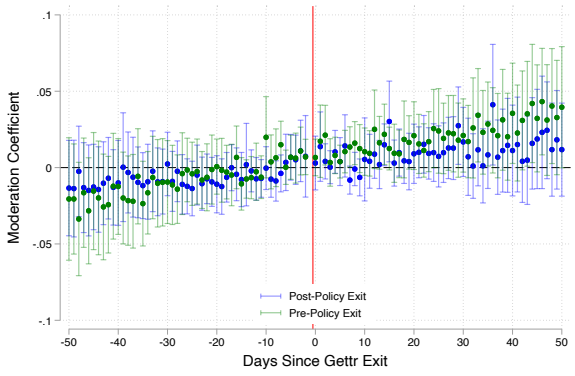
(c) Gab - After

Note: This figure shows the probability that users mentioned content moderation in the 30 days prior to their disengagement from the platform (i.e., last post date), plotted by deciles of pre-treatment threatening content. Panels (a)–(c) report estimates for the pre-policy period (users who exited before July 27, 2022), while Panels (d)–(f) show the immediate post-policy period (users active after July 2022 but exiting before November). For each platform, I regress a binary indicator for whether the user ever mentioned moderation in their final 30 days on pre-treatment threat deciles, controlling for users' average post frequency and timing relative to the policy announcement. 95% confidence intervals shown.

Figure B.16: Event study of Gab discourse before and after Gettr exit, by policy period



(a) Threatening content

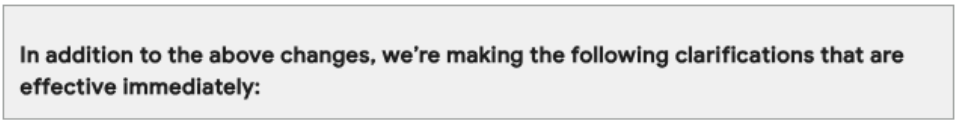


(b) Moderation-related discourse

Note: Each panel plots event study estimates of Gab user behaviour in the days surrounding their last post on Gettr. Treated users (blue) are those who exited Gettr after Google's July 2022 policy change; control users (green) exited Gettr prior to the policy. Outcomes are (a) the threat score of Gab posts, and (b) the likelihood of moderation-related language. Estimates are from separate regressions of the outcome variable on event-time dummies (days relative to Gettr exit), absorbed by user and date, with standard errors clustered at the user level. The baseline period is day −1.

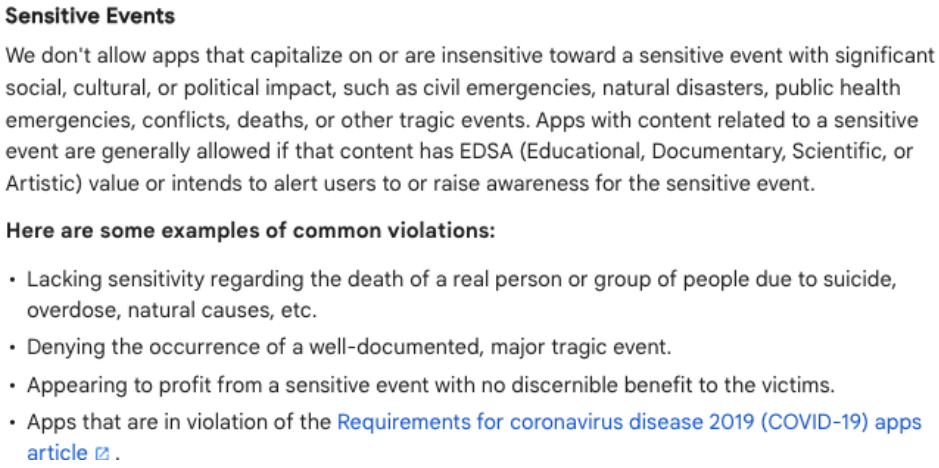
C Other Content

Figure C.1: Google announcement - Sensitive Events



Note: Excerpt from Google developer policy update on July 27,2022. See [Archive]

Figure C.2: Google announcement - Sensitive Events



Note: Excerpt from Google developer policy update on July 27,2022. Appeared under the heading in Figure C.1. See [Archive]

Figure C.3: Google announcement - User Generates Content Moderation

Effective October 11, 2022

User Generated Content

User-generated content (UGC) is content that users contribute to an app, and which is visible to or accessible by at least a subset of the app's users.

Apps that contain or feature UGC, including apps which are specialized browsers or clients to direct users to a UGC platform, must implement robust, effective, and ongoing UGC moderation that:

- Requires users accept the app's terms of use and/or user policy before users can create or upload UGC;
- Defines objectionable content and behaviors (in a way that complies with Google Play Developer Program Policies), and prohibits them in the app's terms of use or user policies;
- Conducts UGC moderation, as is reasonable and consistent with the type of UGC hosted by the app;
  - In the case of augmented reality (AR) apps, UGC moderation (including the in-app reporting system) must account for both objectionable AR UGC (e.g., a sexually explicit AR image) and sensitive AR anchoring location (e.g., AR content anchored to a restricted area, such as a military base, or a private property where AR anchoring may cause issues for the property owner).
- Provides an in-app system for reporting objectionable UGC and users, and takes action against that UGC and/or user where appropriate;
- Provides an in-app system for blocking UGC and users;
- Provides safeguards to prevent in-app monetization from encouraging objectionable user behavior.

Note: Excerpt from Google developer policy update on April 6,2022. See [Archive]

Table C.1: Major Infrastructure-Based Enforcement Actions Against Alt-Tech and Fringe Platforms

Platform	**Year**	**Infrastructure Provider**	**Type of Infrastructure**	**Reason for Enforcement**	**Outcome**
Gab	2017	Google Play Store	App Store	Violated hate speech policy; insufficient moderation	Removed from Play Store; later dropped lawsuit against Google
Gab	2018	Joyent, GoDaddy, PayPal/Stripe	Hosting, Domain, Payment Processor	Links to extremist violence (Pittsburgh synagogue shooting)	Hosting and payment services suspended; domain removed; platform migrated
Parler	2021	Apple and Google App Stores	App Store	Failure to moderate violent content post-Jan 6 Capitol riot	Removed from both app stores
Parler	2021	Amazon Web Services	Hosting Provider	Hosting of violent content; inadequate content moderation systems	Hosting suspended; site went offline temporarily
TruthSocial	2022	Google Play Store	App Store	Failure to moderate violent content; screenshots of inciting posts shared by Google	Initially rejected; approved after improving moderation systems
8chan (later 8kun)	2019	Cloudflare, Voxility	CDN / DDoS Protection	Hosting manifestos related to mass shootings	DDoS protection revoked; site went offline temporarily

Continued on next page

Table C.1 – continued from previous page

Platform	Year	Infrastructure Provider	Type of Infrastructure	Reason for Enforcement	Outcome
The Daily Stormer	2017	GoDaddy, Google Domains, Cloudflare	Domain, CDN / DDoS Protection	Hate speech and involvement in Charlottesville rally	Domain and CDN services revoked; site forced to migrate
Kiwi Farms	2022	Cloudflare	CDN / DDoS Protection	Targeted harassment and doxxing campaigns	DDoS protection revoked; site went offline temporarily
Wimkin	2021	Apple and Google App Stores	App Store	Violent content and failure to moderate post-Jan 6	Removed from app stores
Al-Manar	2012	Apple and Google App Stores	App Store	Content promoting hate and ties to designated terrorist organisations	Removed from add stores

Table C.2: Examples of Dictionary Terms by Topic

Topic	**Representative Patterns / Terms**
Election Fraud	`ballot`, `fraud`, `2000 mules`, `stop the steal`, any combination of `steal + election`
Anti-Vaccine	`hoax`, `plandemic`, `scamdemic`, `Bill Gates`, `Pfizer`, `vax`
QAnon	`WWG1WGA`, `The storm is coming`, `Q drop`, `Deep state`, `Adrenochrome`, `Save the children`
Climate Change Denial	`climate hoax`, `geoengineering`, `Agenda 21`, `green tyranny`, `chemtrails`
January 6 Commentary	`J6 patriot`, `Free the J6`, `Ray Epps`, `False flag`, `Political prisoners`, `Ashli Babbitt`

Note: Full regular expression lists are available upon request and included in the replication package.

Figure C.4: House Oversight Committee to Gettr - August 19, 2022

CAROLYN B. MALONEY, NEW YORK
CHAIRWOMAN

ONE HUNDRED SEVENTEENTH CONGRESS

JAMES COMER, KENTUCKY
RANKING MINORITY MEMBER

Congress of the United States
House of Representatives
COMMITTEE ON OVERSIGHT AND REFORM
2157 RAYBURN HOUSE OFFICE BUILDING
WASHINGTON, DC 20515–6143
MAJORITY (202) 225–5051
MINORITY (202) 225–5074
https://oversight.house.gov

August 19, 2022

Mr. Jason Miller
Chief Executive Officer
Gettr
3 Columbus Circle, 20th Floor
New York, NY 10019

Dear Mr. Miller:

We write regarding your company's response to the surge of online threats against law enforcement following the execution of a court-authorized search warrant by the Federal Bureau of Investigation (FBI) at former President Trump's Mar-a-Lago Club. We are concerned that reckless statements by the former President and Republican Members of Congress have unleashed a flood of violent threats on social media that have already led to at least one death and pose a danger to law enforcement officers across the United States. We urge you to take immediate action to address any threats of violence against law enforcement that appear on your company's platforms.

On August 8, 2022, the FBI conducted a search at Mr. Trump's private club in connection with the former President's potential violations of the Espionage Act, the Presidential Records Act, and other federal laws. Following the search, former President Trump and House Republicans lashed out at the FBI and law enforcement more generally. While FBI agents were executing the court-approved search warrant, former President Trump released a statement claiming his home was "under siege, raided, and occupied by a large group of FBI agents."[1] He later claimed, without evidence, that the FBI may have planted evidence.[2] House Minority Leader Kevin McCarthy accused the Department of Justice of being "weaponized" against former President Trump. Minority Whip Steve Scalise suggested that the FBI agents executing the warrant had gone "rogue."[3] Republican Members of Congress wrote on Twitter, "We must

[1] *FBI Executes Search Warrant at Trump's Mar-a-Lago in Document Investigation*, CNN (Aug. 9, 2022) (online at www.cnn.com/2022/08/08/politics/mar-a-lago-search-warrant-fbi-donald-trump/index.html).

[2] Donald J. Trump (@realDonaldTrump), Truth Social (Aug. 10, 2022) (online at https://truthsocial.com/@realDonaldTrump/posts/108798211943189544).

[3] See *McCarthy Threatens to Probe Garland After Trump FBI Raid*, The Hill (Aug. 8, 2022) (online at https://thehill.com/policy/national-security/3593582-mccarthy-threatens-to-probe-garland-after-trump-fbi-raid/); *Asked About Threats to FBI, Scalise Claims Without Evidence that Agents Went 'Rogue'*, Washington Post (Aug. 11, 2022) (online at www.washingtonpost.com/politics/2022/08/11/trump-fbi-search-scalise-republicans/).

Note: This is a letter sent to Gettr CEO Jason Miller by the House Oversight Committee on August 19, 2022. Identical versions were also sent to the CEO's of Facebook, Twitter, Truth Social, Gab, Rumble, Telegram, and TikTok.

Table C.3: Timeline of Apple App Store Review Policy Updates in 2022

Date	Summary of Policy Changes
2022-04-05	Apple permits "reader" apps (e.g., for books, audio, video) to include an in-app link to a developer-controlled website for account creation and management. This relaxation of 3.1.3(a) is formalized via the *External Link Account Entitlement*. However, such apps still cannot promote alternative payment methods within the app.
2022-06-06	Guidelines updated ahead of upcoming OS releases: • Reinforced rules on overtly sexual content (1.1.4) and background service use (2.5.4). • Clarified that law enforcement involvement is required for apps reporting criminal activity. • Strengthened rules against use of health-related APIs (e.g., HealthKit) for advertising or data mining. • Emphasized that in-app purchases cannot be used to buy credit for real-money gaming (5.3.3). • Required clear developer contact information in all apps, especially those used in classrooms. • Reiterated restrictions on using Apple interfaces, emoji, or music assets without authorization (5.2.5).

Date	Summary of Policy Changes
2022-10-25	Major clarification and expansion of content governance and monetization rules: • New prohibition on apps exploiting **sensitive events** (e.g., conflicts, epidemics) for profit (1.1.7). • Expanded 1.1.4 to explicitly mention human trafficking and exploitation as grounds for rejection. • NFTs: Apps may use in-app purchase to sell NFTs and related services (minting, listing, transferring), but cannot link to external marketplaces. • Cryptocurrencies: Apps may enable exchange-based transactions only in licensed jurisdictions. • Display advertising must remain within the main app binary and avoid sensitive targeting (2.5.18). • Matter-supported apps must use Apple's Matter SDK or certified alternatives. • Clarified demo access expectations: developers must offer a working demo account or a full-feature demo mode. • Specified that in-app purchases are required to unlock functionality (no external QR codes, crypto wallets, etc.). • New exemption for advertising campaign management apps from in-app purchase requirements (3.1.3(g)).