Can hate speech be banned online? The effects of shutting down toxic forums on Reddit*

Adam Di Lizia† Lily Shevchenko‡

October 20, 2026

Click here for the latest version

Abstract

Is deplatforming effective in reducing levels of hate speech on social media? To answer this question we study a policy change on Reddit which led to a simultaneous ban of thousands of forums containing hateful content, but not the users of those forums. We use data on the near universe of comments left on Reddit to examine the impact of the ban on user behaviour in a differences-in-differences design. We find that the most active users of banned subreddits comment more after the policy change and substitute to new forums in the weeks after the ban. The increase in activity persists in the long run, but is not associated with higher levels of hate speech: instead, the comments left by affected users outside banned subreddits contain 20% fewer instances of slurs. We do not find evidence that the policy leads to lower quality of engagement, negative spillover effects or re-creation of banned forums elsewhere on the platform. Overall, the results suggest that community-level moderation can be effective in combating hate speech without lowering user engagement, and thus can be aligned with platforms' incentives.

*We are grateful to Manuel Bagues, Sascha Becker, Dan Bernhardt, Apurav Bhatiya, Mirko Draca, Dean Eckles, Ruben Enikolopov, Thiemo Fetzer, Marita Freimane, Andrew Funck, Thomas Hill, Raphael Jimenez-Duran, Ro'ee Levy, Agustina Martinez, Mike McRae, Juan Morales, Maria Petrova, Marina Rizzi, Ananya Sen, Mateusz Stalinski, Gerhard Toews, Ao Wang, Ekaterina Zhuravskaya, the participants of the 2025 NICEP conference, 2025 Warwick PhD Conference, 2025 PhD Conference in HEC, 2025 PPE Conference in King's College London, 2025 CODE@MIT, 2026 Doctorissimes Conference, 2026 Natural Language in Economics Workshop, 2026 Applied Young Economist Workshop, 2026 QPE Early Career Workshop, 2026 EAYE Annual Meeting, 2026 South West Economics PhD Conference, 2026 POLECONUK Annual Conference, 2026 CESifo Workshop on Digital Platforms, 2026 RES Annual Conference, as well as the participants of internal seminars in Warwick and LSE for their comments and suggestions. This study was approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick. Our work has benefited from access to the Scientific Computing Research Technology Platform at the University of Warwick. We acknowledge the support of the Centre for Advantage in the Global Economy (CAGE), Adam Di Lizia additionally thanks the ESRC for funding this research agenda. All remaining errors are our own.

†Department of Economics, University of Warwick. Email: adam.di-lizia@warwick.ac.uk

‡Department of Economics, University of Warwick. Email: liliia.shevchenko@warwick.ac.uk

1 Introduction

Consumers, platforms, and policymakers are increasingly aware of the social and economic effects associated with proliferation of hate speech online. In addition to affecting those exposed to it directly (Braghieri et al. 2022), hateful content contributes to misinformation (Ershov and Morales 2024), polarization (Allcott and Gentzkow 2017) and hate crime (Cao et al. 2023; Müller and Schwarz 2020; Müller and Schwarz 2023). Since toxicity online can elicit higher levels of engagement, algorithmic recommendation systems of social media can exacerbate the problem by amplifying and spreading hate speech (Beknazar-Yuzbashev et al. 2025; Guess et al. 2023; Levy 2021).

In the last decade, most social media platforms have been introducing progressively stricter moderation policies to curb hateful content, often spurred by legislative changes. Mounting empirical evidence suggests that the various types of such interventions tend to successfully reduce levels of toxicity and hate speech, but the effect sizes are small (Aridor et al. 2024). Despite these measures, hate speech and toxicity are still widely prevalent online: recent analysis by Ofcom suggests that 77% of internet users in the UK have encountered hateful content in the last year, with more than half reporting weekly exposure (Ofcom 2026).

At the same time, little is known about the direct effects of moderation on users that it targets, *suppliers of toxicity and hate speech*. If deplatforming alienates extreme users, then moderation interventions can push them to engage in censorship evasion and migrate into smaller, less-regulated digital spaces, while still reducing the observed incidence of hateful conduct on a platform.

In this paper, we study the effects of moderation on users of hateful forums by leveraging a policy change on Reddit, a popular social media platform with over 700 million active monthly users and one of the most visited websites in the world.[1] In June 2020, Reddit changed its terms of use to prohibit hateful content and immediately banned a large number of forums violating the new rules – but not their users. This exogenous intervention enables us to study migration of displaced (treated) users and differences in the way they interact with others on Reddit.

Our empirical analysis is based on a dataset of over 2.1 billion comments left by 29 million users of Reddit in 2020 available via the `Pushshift` database (Baumgartner et al. 2020). We identify users who left at least one comment in subsequently banned forums, resulting in a set of approximately 190,000 treated accounts. We estimate a differences-in-differences specification exploiting temporal variation around the ban and cross-sectional variation in individual exposure to the policy change.

Our first set of results concerns migration of displaced users across Reddit in the aftermath of the ban. For that purpose, we construct a panel of the number of comments

[1] As of 2025, Reddit ranks as 10$^{\text{th}}$ most popular social media platform and 6$^{\text{th}}$ most visited website in the world (Duong 2025a, Duong 2025b).

each user left in a given week. We find that users most active in banned forums prior to the policy change leave, on average, 10% more comments across the rest of the platform afterwards. The results are robust to alternative specifications and in triple-difference design using a placebo set of non-banned forums of similar size and patterns of user activity. The increase in activity is sustained in the long run and is associated with higher activity levels in previously visited forums, as well as search of new forums in the weeks after the ban. We do not find evidence that banned communities are re-created elsewhere on the platform or that the ban is associated with substantial platform exit. Overall, our results suggest that the shutdown of toxic forums led to substitution to other forums on Reddit, but not in a coordinated manner.

To quantify the effects of the ban on toxicity of affected users we measure incidence of hateful language for every comment in the dataset using a purpose-built dictionary of hate speech. In a comment-level estimation, we document that the ban led to an approximately 20% average reduction in number of slurs used by the treated users *outside banned forums*. We show that the effect is not explained by changes in moderation policies within forums or exit of highly toxic users and is robust to alternative measures of hate speech. We document that levels of toxicity and hate speech use on Reddit are primarily driven by community norms as opposed to inherent characteristics of individual users. Our results reflect a decrease in hate speech incidence relative to other users in the same conversation types as opposed to selection into discussions with lower local level of toxicity. Reduction in slur use has beneficial spillovers on non-treated and other treated users in the same discussions through reduction in contagion. At the same time, we do not find negative effects on quality of engagement either of treated users or those interacting with them, as proxied by text length, number of links used, probability to quote other users or number of replies to a comment.

Lastly, we document that the overall results conceal important heterogeneity in response to the policy change by the political leaning of treated users. The increase in activity is driven by most active users of banned left-wing and transphobic forums. At the same time, reduction in use of hate speech is almost entirely due to users of banned right-wing forums. We suggest that both results are primarily driven by relative distance in community norms between subgroups of treated users and the rest of the platform: left-wing and transphobic users are less dissimilar from Reddit as a whole, find it easier to substitute to alternative forums and do not have to adjust their language as drastically.

Overall, our results point to the conclusion that moderation targeting hateful forums on Reddit is effective in curbing hate speech beyond mechanical effects due to censoring: it lowers the use of slurs by treated users outside banned forums and leads to additional beneficial spillover effects on other users in the same conversations. However, heterogeneity in language adjustment and in degree of reallocation of displaced users across Reddit suggest that *characteristics of the group of affected users matter.* Specifically, substitution

costs vary by partisan orientation and contribute both to the effectiveness of the policy in reducing hate speech use and to the ability of the platform to retain moderated users after the intervention.

Related literature Our primary contribution is to economic literature on the effectiveness of content moderation in reducing online toxicity. Several studies show that regulatory changes have been largely successful in inducing platforms to strengthen moderation. Jiménez-Durán et al. (2023) leverage a change in German legislation regarding fines on social media platforms that host hateful content (NetzDG) and find that the law induces Twitter to moderate toxicity on the platform. McRae (2025) shows that upstream infrastructure providers can also have meaningful impact on platform's moderating decisions: the paper documents that the prospect of losing access to Google's Play Store led to a decrease in threatening content and increased moderation of politically sensitive narratives on Gettr and Truth Social.

But the effectiveness of moderation interventions *themselves* in curbing toxicity and hate speech appears to be limited. The same paper analysing the effect of NetzDG finds a small associated decrease in toxicity although it is unclear to which degree it can be attributed to censored content itself as opposed to change in user behaviour (Jiménez-Durán et al. 2023). Similarly, Jiménez-Durán (2023) finds in experimental setting that additional reporting of comments containing slurs on Twitter does not change the toxicity of the reported users in the 5 months after the intervention. Horta Ribeiro et al. (2023a) look at posts just below and above a toxicity threshold for moderation on Facebook and report that users whose comments have been removed are slightly less likely to be in breach of community guidelines for a short time after the policy.

Moderation interventions that disrupt the media "consumption basket" of a user appear to be more effective. The intuition that we propose is that users' behaviour online is shaped in large part by what they are exposed to – through contagion, persuasion, and by shaping perceptions of the social norm. Individual-level interventions might have limited effects on targeted users since they do not affect any of those pathways. On the contrary, in line with this argument, Beknazar-Yuzbashev et al. (2025) report that reducing exposure to *toxic posts* in algorithmic feeds can lower production of toxic content. Thomas and Wahedi (2023) investigate a staggered ban of hate organizations' leaders on *Facebook* and find that deplatforming *hateful users* can also lower the levels of hateful content consumed and produced by audiences of these groups, as well as result in gradual reduction of network strength of hate organizations. A similar pattern but for a different set of outcomes is reflected in recent work by Gauthier et al. (2026) on the impact of algorithmic recommendation on misinformation and adherence to conservative ideology. The paper finds that while switching on algorithmic feed on Twitter/X increases following of conservative accounts and changes beliefs on politically salient topics, switching the algorithm off does

not have symmetrically opposite effects. The authors explain the result by network persistence: once users have started following conservative accounts suggested by the algorithm, *exposure* to this type of content is not disrupted by the switch to chronological feed.

This paper furthers this line of research by analysing a rare case of a targeted ban of a large number of forums promoting hateful content on Reddit. The ban drastically reduced individual exposure of the users of banned forums to hateful content and simultaneously decreased their connection to the network of other treated users. We document that negative shock to exposure to hate speech can lower toxicity, even for users we tend to think of as suppliers of toxic content.

We report a large (approximately 20%) average decrease in hate speech use on Reddit due to the ban. Our argument is that, at least in this setting, use of hate speech is more dependent on contagion than on the inherent user type. A drastic reduction in contagion generates disproportionately large effects for users with high *baseline* level of hate speech use, further amplified by positive spillover effects. Reddit's organization as a forum-based platform contributes to the effectiveness of the ban: individual users of the same forum who have not communicated directly cannot easily find each other after the policy change and the network of users of a banned forum cannot be easily "transplanted" elsewhere on the platform.

Second, we contribute to literature on platform migration of politically active users by drawing the attention to differences across political spectrum. Research points out that moderation interventions can generate backlash and/or migration to alternative, less regulated platforms, jeopardizing the effectiveness of moderation when we consider impacts beyond a single platform. This scenario appears especially likely if moderation is perceived as unjustified or unfair by platform users. In the case of moderation of hate speech, this is particularly likely, because there is a tight link between extreme political opinions and specific classes of hate speech. Rizzi (2024) finds that Twitter's 2020 moderation policy aimed at curbing racist language was effective in reducing hate speech use, but led to increase in activity and toxicity on Parler. Following Parler's own takedown in 2021, literature documents coordinated migration to Telegram (Agarwal et al. 2022) and increase in activity on other alternative platforms (Horta Ribeiro et al. 2023b). On the other hand, McRae (2025) shows that deplatforming on alternative media platforms does not necessarily generate significant backlash and coordinated migration when the intervention is not highly publicised.

We show that the ban of forums promoting hate speech on Reddit similarly did not lead to substantial exit from the platform or re-creation of banned forums elsewhere on Reddit, even though the ban itself was highly publicised. We propose two main mechanisms. First, backlash to moderation is highly dependent on ability of users to coordinate. Community-level moderation on Reddit is effective in breaking up toxic communities, since connections between individual users are much weaker than on Facebook or X (formerly

Twitter). Second, platform exit requires existence of suitable substitutes. Since the ban on Reddit affected both extreme left and extreme right, we can provide estimates across the political spectrum (previous work has been largely focused on the right-wing). We find evidence of substitution to the rest of Reddit for users of left-wing and trans-exclusionary forums, but not for the alt-right. All of the alternative platforms for which user migration has been documented previously (Parler, Gab, 4chan) are predominantly right-wing and are a natural coordination point neither for left-wing nor for transphobic users. Lack of substitutes is likely part of the reason they stay on Reddit.

Regardless of the actual mechanisms at play, our results highlight that user response in the aftermath of a moderation intervention can be driven by a variety of context-dependent factors. Importantly, as is the case in our paper, that selection into specific policy response can vary by partisan affiliation, implying that the same policy can have *unintentionally uneven* effects depending on the subgroup. This is in line with work by Ershov and Morales (2024) showing that an intervention on Twitter aimed at curbing misinformation by making it harder to share content has had disproportional effect on left-wing outlets due to higher sensitivity to sharing cost. Here we document partisan differences in response to moderation. Overall, our results point to the conclusion that the effects of moderation reported in the literature for far-right content not be immediately generalizable to other types of extreme behaviour online.

We also relate to the more general academic discussion on the determinants of social media use, as well as the tension between user engagement and welfare. Notably, Allcott et al. (2020) suggest that addictive properties of social media lead users to spend more time on Facebook than would be optimal. Beknazar-Yuzbashev et al. (2025) show in experimental context that toxic content, while promoting engagement and thus being beneficial for platforms relying on ad shows for revenues, is not necessarily welfare maximising, and that lowering exposure to toxicity can reduce user engagement within platform and lead to substitution to more toxic platforms. The policy change on Reddit is a rare opportunity to study patterns of substitution in response to an exogenous shock to the "media consumption basket" of treated users in observational data. We find evidence that users of banned forums increase activity in forums with higher average incidence of toxicity and hate speech, suggesting that substitution to the rest of Reddit that we observe after the policy might have similar mechanisms to off-platform substitution observed by Beknazar-Yuzbashev et al. (2025).

At the same time, the reductions in level of hate speech use that we observe are not associated with significantly lower levels of engagement or a decrease in engagement quality of treated users and users replying to them. Similarly to Beknazar-Yuzbashev et al. (2025), we argue that the relationship between hate speech and levels of engagement might not be the same as the relationship between toxicity and engagement. Thus, policies curbing use of hate speech might not face the same issue of going against platform interests as policies

reducing exposure to toxicity do.[2]

Finally, we contribute to substantial interdisciplinary literature studying community norms and moderation on Reddit. Overall, this line of research clearly indicates lack of consensus on the effects of various moderation interventions within Reddit. Chandrasekharan et al. (2017) study the effectiveness of the 2015 ban of two forums promoting violence and harassment on activity and use of hate speech by affected users. The paper documents large decreases in use of hate speech and migration of users to other forums. Cima et al. (2024) and Cima et al. (2025) look at the impacts of the same 2020 policy change that we study on platform activity and toxicity for users of the 15 largest banned forums. Both papers report that a large fraction of affected users leave Reddit altogether and that there was a small overall reduction in toxicity with some evidence of alienation and increased toxicity for a small fraction of the users. Trujillo and Cresci (2022) focus on the effects of subsequent restrictions of r/The_Donald on the behaviour of its users within the forum and on the rest of the platform: unlike previous work, the paper reports a decrease in activity and an increase in toxicity of the members, as well as negative effect on the quality of information shared. All of this literature stresses heterogeneity in response depending on the type of moderation intervention and warns of potential repercussions due to alienation of most active users, movement to alternative platforms with lower levels of moderation, and negative spillovers onto the rest of Reddit.

We advance this line of research in three primary directions. First, we characterise the effects of the ban for a wider range of outcomes and provide empirical evidence on some of the mechanisms underlying the observed effects. Second, we draw particular attention to heterogeneity not only due to the degree of user engagement in a banned forum, but also due to to the partizan differences across banned forums. Third, we put higher emphasis on causal analysis of the policy change. Most of our results are quantitatively different to those reported in previous work, have different interpretation and implications for platforms and policy makers.

The rest of this paper is organised as follows. Section 2 provides background on the general approach to moderation on Reddit and specifics of the policy change. Section 3 describes the data sources and construction. Section 4 discusses estimation strategy and the effects of the ban on migration of users across the platform. Section 5 discusses the estimated impact on the use of hate speech. Section 6, elaborates on the relation of the results to previous literature and policy implications. Section 7 concludes.

[2]Our results here can only be interpreted as suggestive, since we do not observe page views or activity outside Reddit.

2 Background

Reddit is a primarily English-language social media platform founded in 2005. As of 2026, Reddit has approximately 130 million daily active users; in 2021, at the end of our sample period, that number was approximately 54 million (Duong 2026a). Around half of Reddit's traffic comes from the United States (Duong 2026b); users from Canada, United Kingdom, Australia, India, Germany, France and Brazil comprise the largest share of the remaining traffic (Statista 2024). Similarly to most other social media platforms, Reddit is primarily monetized through advert shows. At the time of the policy change we study in this paper, the company was privately owned.

Unlike on most other major social media platforms, activity on Reddit is centred not on personal pages of users and individual news feeds, but on forums, aptly called *subreddits*. *Subreddits* are essentially boards of posts tied to a specific topic, such as `r/swimming` or `r/politics` (see Figure D.1 for an example). The primary mode of interaction on Reddit is participation in *subreddits*, where users can post (text, images or videos), comment, and react (with "upvote" or "downvote") to posts and comments of other users. The sum of upvotes and downvotes is displayed as a score, and posts, as well as comments, can be sorted by the score or in chronological order (Figure D.2 provides example of the interface). Additionally, users can create private group chats or message other users directly.

Users can visit individual *subreddits* or browse the *front page* of the website displaying most popular posts across the platform. The front page tends to reflect general trends of the Internet: discussion of news, sport events and media, trending memes and videos, et cetera. In most cases, anyone can browse content of a forum, but only logged-in users can leave posts and comments. Forums themselves can impose additional restrictions (for example, being member of a *subreddit* for 30 days to be able to post) to prevent unwanted activity: bots, trolling, and spam.

Content moderation on Reddit *Subreddits* are run not by the platform directly, but by volunteer members called *moderators*. Moderators set the rules for behaviour on the forum and have the power to delete posts and ban users violating these rules. Additionally, individual users can report content or simply downvote a piece of content, contributing to its lower visibility and proliferation rate. In general, Reddit has been consistently promoting strong *internal* moderation of forums, as opposed to platform-wide policies, since "the only thing that scales with users is users themselves" (Huffman 2019). In company's opinion, the resulting "federated" system of governance can maintain locally appropriate rules of communication as the platform grows not through costly centralized monitoring, but through creation of *local norms of interaction* enforced by communities themselves (Huffman 2019).[3]

[3]As a result, forum rules can be very specific to a *subreddit*: for example, `r/politics` does not allow users to post links to news outlets outside the published set of trustworthy sources.

On platform's side content moderation is done by *admins*, largely responding to automatically flagged content violating Reddit's content policy and removal requests submitted by users. According to the transparency reports published by Reddit, in 2020, "content manipulation" (activity trying to artificially promote content) accounted for 99.7% of all content removed by the platform itself. Vast majority of non-spam comments are removed by local moderators aided by AutoMod – automatic system that can be configured to forum-specific rules.

Admins can additionally impose sanctions on individual user accounts, ranging from warnings to short-term (up to a week) suspensions to permanent bans. In 2020, only 2.5% of permanent bans were issued due to a violation of content policy, with 85% of bans citing spam as the reason for suspension and 13.5% citing evasion of previous account-level bans. Every year Reddit removes content and bans users for other reasons, mostly due to requests from government agencies, private entities and due to copyright violations, but these constitute a negligible proportion of total removals on the platform.

Historically, moderation interventions targeting whole forums have been rare. Forum-level sanctions include quarantine and ban, and are always imposed by admins. Quarantine is imposed in the rare cases when the forum does not violate content policy but its content can be offensive to general public. Quarantined *subreddits* are not limited in their normal activity, but are restricted in reach: they are excluded from search and recommendation results, do not appear on personal feed unless a user is subscribed to the forum, and display a content warning for a new user trying to view the content. Additionally, thousands of forums are banned every year, upwards of 95% of them due to lack of moderation, ban evasion and spam. Although the exact numbers are not available, qualitative evidence suggests that the vast majority of banned forums are very small.

Individual bans of larger forums, when they are not imposed due to the prevalence of illegal content, are highly publicized and usually generate controversy regarding implications on free speech. Prior to 2020 these bans are rare – largely because Reddit *did not formally prohibit* non-illegal content in its Content Policy for most of its history, meaning that there were no formal grounds for such bans. In 2015, in the only major relevant content policy overhaul prior to our sample period, Reddit prohibited violence and harassment, introduced the concept of "quarantining" and banned a slate of communities "making Reddit worse" (Huffman 2015). Since then, several other prominent forums have been banned citing similar reasons: in 2017 `r/incels` was banned for promoting violence against women, `r/altright` was shut down for publishing private information about a person who attacked a prominent right-wing public figure. But, according to Reddit's own statements, until 2020 the platform did not have specific rules against hate speech, because Reddit "struggled to create a definition of hate and racism that we could defend and enforce at our scale" (Huffman 2020a).

Policy change The platform was forced to take a more definitive stance on hate speech in June 2020, in the wake of George Floyd's murder and mounting pressure to curb the role of Reddit in spreading hateful and violent rhetoric. On 8th June 2020 hundreds of moderators called out Reddit's CEO Steve Huffman for continuous neglect of the issue and urged Reddit to be more proactive in banning hateful forums and their users (r/AgainstHateSubreddits 2020).

As a result, on June 29th 2020 Reddit changed its content policy to clarify, among other things, definitions of hate and state that "communities and users that promote hate based on identity or vulnerability will be banned". Simultaneously, the platform banned approximately 2000 *subreddits* violating the rules, the vast majority of them inactive at the time of the policy change (Huffman 2020b). The move received mixed feedback with many moderators and outlets stating that the change is "reactive" and a "PR trick" (Cite Wired). The platform was also criticized for lack of transparency (neither the full list of banned forums nor the exact reasoning behind their ban was ever published), lack of bans for some specific types of hate speech (notably, extreme Hindu nationalism), as well as lack of concrete actions to make moderators' work easier in the future.

The ban was not exclusive to forums promoting white nationalist ideologies: Reddit attempted an "even-handed approach", likely to signal that the moderation intervention is not politically motivated. *Subreddits* banned in June 2020 for spreading hateful and violent content can be classified in three major groups: right-wing forums containing xenophobic and transphobic content as well as highly offensive humour; "dirtbag left" political *subreddits* extensively using offensive language and violent rhetoric, and *subreddits* containing extreme misogynistic content, largely of sexualized nature (for more detailed characterisation of banned forums see Section 3). In this paper, we focus attention on different subtypes of right- and left-wing content, since we expect the suppliers of misogynistic sexual content to have a completely different mode of interaction within the platform. Still, the policy change provides a rare opportunity to study effects of the *same moderation intervention across a wide spectrum of hateful content*. Recent literature has largely been focused on the effect of moderation on US-focused alt-right content, while other types of hate speech received relatively limited attention. Our empirical setting allows us to see whether moderation effects are actually symmetric across those groups.

3 Data

The main data source for our analysis is `Pushshift` (Baumgartner et al. 2020): a publicly available archive of the near universe of comments and posts on Reddit from the creation of the platform in 2005 till 2023, when the official Reddit API stopped being accessible to general public.

`Pushshift` data contains text of each comment, the name of the user who left the comment, *subreddit* (forum) where it was posted, creation timestamp, and identifiers allowing us to pinpoint under which post the comment was left and the relative position of the comment within the discussion thread. We also observe reactions to the comment in the form of score (sum of "upvotes" and "downvotes" left by other users on the comments), as well as some other fields.

Given that large fraction of posts are links or media with little to no text, and that it is the comments *replying* to the posts that constitute the bulk (upwards of 85%) of the public activity on Reddit, we focus on comments as main unit of the analysis. For the majority of exercises, we consider all comments left on Reddit in year 2020 – that gives us an approximately six month period before and after the policy change.

Data processing Data processing for this analysis was limited to variable type conversions, anonymization of user accounts and transforming the text of comments into format suitable for natural language processing (NLP) tools. For that purpose, we flag and separate quotes, URLs, tables, code blocks and other non-textual content. We then normalize text syntax, formatting and encoding. These steps ensure that variables such as text length are precise and give us access to additional information, such as number of quotes per comment or the set of links used. To obtain tokenized and lemmatized text used for dictionary lookups, we process cleaned text using `SpaCy` (Honnibal et al. 2020).[4] Additionally, we predict the language of each comment using `langdetect`[5] library and exclude non-English language comments from exercises including text-based measures.

All exercises additionally exclude comments left by bots. We identify bots via standard techniques filtering users who comment in two different *subreddits* at exactly the same timestamp, accounts leaving large volumes of approximately identical text, as well as through manual inspection of accounts on the upper tail of total activity distribution. We do not expect these observations to have any impact on the estimates, but removing bots helps make the analysis computationally feasible: it results in the removal of almost 63 million comments (3% of total dataset).

[4] We use the lightweight `en_core_web_sm` model since it performs best within the class of models that are computationally feasible for us to use.

[5] https://github.com/Mimino666/langdetect

Missing data One limitation of the data is that comments left in subsequently banned *subreddits* from second week of April until the ban in June 2020 are not available. The reason is that `Pushshift` saves all comments every 3 days but with a lag of 2 to 3 months. Reddit restricted API access to historical data from banned forums at the time of the policy change, and lagged `Pushshift` data stops at the same time. The advantage of the delay is that, since most activity in a thread happens within short period of time after its creation, we likely have near complete data on all comments in a post as well as close to "final" snapshots on such variables as scores and awards. Lack of data from banned forums for April-June 2020 is not a serious issue for the empirical results presented in this paper for two main reasons.

Firstly, as is discussed in more detail in Section 4.1, we *purposefully avoid including comments from banned subreddits* in our analysis to eliminate mechanical effects of the policy on the outcomes of interest, thus comments from banned forums do not enter the analysis directly. Secondly, although lack of data for April-June 2020 does lead us to incorrectly assign treated users who only appear in banned forums in that timeframe to the control, this misattribution cannot have any meaningful impact on the estimates due to the large size of the control group. We also expect that users only commenting in banned forums from April 2020 are much more likely to be "incidental" – engaging in banned forums because of the contemporaneous political events, while our main focus is on the *core* users of banned forums, for whom the exposure to the policy change is the greatest.

Naturally, `Pushshift` database also does not include activity in private chats or private *subreddits*- forums content of which is restricted to members only, so the analysis in this paper reflects the "public" side of the activity on the platform. We do not have official data on the proportion of private forums for 2020, but estimates from more recent datasets put the number at around 6% of total forums. The effective share of private forums from total activity on Reddit is much smaller, since member-only forums are by definition small. Research analysing completeness of `Pushshift` data within 2005-2016 timeframe suggests that the dataset is missing at most 0.043% of total comments which is the upper bound on the total number of comments left in private and banned forums in that timeframe. Additionally, we exclude about 10.5% of comments missing user identifier: these are removed comments and comments left by deleted accounts. Unlike observations from banned forums and private subreddits which are "unknown unknowns", deleted comments and user accounts are "known unknowns": we still observe their comment identifier, timestamp and exact location within the discussion chain. We can thus characterise sensitivity of the estimates to missing data.

Banned subreddits and their users The full official list of the approximately 2000 forums banned in June 2020 was never published: the platform disclosed the names of the 10 largest banned *subreddits*, for the rest only the first letters of *subreddit* name and

average daily number of users are available. We use this information to reconstruct the list of all banned *subreddits* with more than 50 average daily users. We do not expect smaller forums to significantly influence the results since the total sum of their average daily user counts does not exceed that of the single largest banned *subreddit*. In this way we identify 112 banned *subreddits*. We define *treated users* as all users who left at least one comment in one of those forums between January and March 2020 (first 14 weeks of the year), the resulting set of users includes 189,950 accounts.

Figure A.1 shows top-20 largest forums by number of comments. Most banned *subreddits* can be classified as either right-wing, left-wing or forums containing extreme misogynistic (and mostly sexual) content. We manually classify each forum into one of those groups. Additionally, since right-wing *subreddits* are clearly more heterogeneous in topic and much less mutually related, we further split right-wing forums into subcategories: alt-right and transphobic forums, *subreddits* containing offensive humour and `rThe_Donald` as a separate category. We assign treated users themselves in the corresponding group if 90% of all their comments within banned forums have been left in the *subreddits* of that category. The results are not sensitive to the specific cut-off and vast majority of treated users never comment in more than one forum type.

We report all estimates for the whole sample of treated users as well as separately for each category excluding misogynistic sexual content. The reason for this decision is that we see that users active in those forums are very distinct from other treated users and users of Reddit in general, and are mostly isolated from the rest of the platform.

There are two main reasons we separate r/The_Donald: firstly, in terms of it's topic it is quite distinct from other "classic" alt-right forums. Secondly, Reddit has targeted r/The_Donald with moderation interventions prior to June 2020, so separating it from the rest of the sample allows us to analyse it in a cleaner manner.

Measurement of hate speech In this paper, we follow the generally accepted definition of hate speech as language targeting groups based on their identity (Kennedy et al. 2020), closely aligned with definition used by Reddit to motivate the ban (Huffman 2020b). At the same time, it is important to be clear that there is no public or academic consensus on which groups are protected under this definition. "Hateful conduct" is a *social construct* in constant flux, its understanding changes over the years and differs across social groups.

The three classic approaches to measurement of hate speech are a) training a purpose-build classifier based on a sample of human-annotated data; b) using an out-of-the box classifier, such as Perspective API and c) matching text to a vocabulary of well-known hateful expressions. All three have their advantages and drawbacks, the choice of vocabulary-based measurement here is motivated by three main considerations.

First, Reddit's policy was focused on reducing hate speech, not toxicity, and the majority of commonly used in economics of media classifiers (Perspective API, Detoxify) do not

differentiate the two exactly. We expect the associated measurement error in our target corpus to be unacceptably high. Second, although classifiers providing better performance in identifying hate speech specifically do exist, it is not computationally feasible for us to produce corresponding estimates for the whole sample of more than 1.8 billion comments or even for a meaningful subsample of the dataset.

Third, dictionary-based approach has clear advantages beyond computational feasibility: we can tailor the measurement to Reddit's particular language and can produce separate estimates for antisemitic, homophobic, transphobic and other slur types. Additionally, the estimation process is transparent, and the results are inherently meaningful and interpretable, even in the presence of measurement error.

To measure incidence of hate speech per comment, we build on the dataset of hate vocabulary by SurgeAI that included hateful and toxic words and expressions, their normalized forms, type (toxicity level or targeted group) and severity score.[6] We manually validate the original dataset containing 1,598 expressions and augment it with data from HurtLex and other public sources. We additionally check that estimates using the dictionary produce meaningful results in comparison with manually labelled data and common classifiers on Measuring Hate Speech dataset by Kennedy et al. (2020).

For every comment in the dataset, we identify each word or n-gram that matches a keyword/key phrase in the dictionary of hate speech either in its raw (tokenized) or lemmatized form. We then manually validate the results to ensure that words that have multiple meanings are not skewing the hate speech estimates per *subreddit* and exclude phrases that would be interpreted as slurs only in rare context-dependent cases (for example, "wandering jew" refers to a plant name in more than 99% of the cases). Validation on *subreddit* level makes the most sense in our setting, since the vocabulary of a forum is largely topic-specific, rather then user-specific. Topic-specific terminology (primarily proper nouns, such as names) accounts for the vast majority of tokens incorrectly labelled as slurs. In addition to manual inspection, we also check that our results are robust to trimming 3 most common keywords for each forum. After identifying keyword matches, we use incidence of toxic language, incidence of slurs overall and incidence by slur type within a comment as the measure of toxicity and hatefulness of a comment, respectively.

Using slur counts has its downsides, mainly that we cannot account for hateful content that is not captured by hateful vocabulary. We expect the measurement error to be almost entirely due to false negatives, meaning that, in general, our estimates of hate speech incidence are likely to be downward biased. A potential concern is correlation between treatment and measurement error: if policy change results in ban evasion, we might be reporting negative effects of slur use when the true effect is null through the relative increase in veiled types of hate speech or uncommon slurs. At the same time, we argue

[6]The data used to be available in the GitHub repository https://github.com/surge-ai/profanity but has been deleted since. We continue using the dataset since it was distributed under the MIT License.

that changes is the use of common slur types, as a measure of *obvious, locally-normalized hateful conduct* is a meaningful object of analysis regardless.

It's worth noting that hate speech is not completely equivalent to hateful content either. Hateful content can take other forms, such as spreading links to misinformation about targeted groups, as one prominent example. Providing evidence on these channels is beyond the scope of this paper, we cite other research looking at the impact on alternative classes of hate-related outcomes in the relevant sections.

4 Effects on user activity on Reddit

4.1 Empirical strategy

Our first objective is to establish whether the ban of hateful forums leads to displacement of activity to the rest of the platform. For this purpose, we construct a panel dataset of the number of comments each user of the platform left in a given week. We then estimate the following difference-in-differences specification:

$$NumComments_{it} = \alpha_i + \tau_t + \beta Post_t \times ShareInBan_i + \epsilon_{it} \quad (1)$$

Where $NumComments_{it}$ is the number of comments left by user i in week t across all non-banned *subreddits*. $Post_t$ is a dummy variable equal to one in the weeks after the ban (weeks 27 to 53) and zero otherwise. $ShareInBan_i$ captures the degree of treated users' *individual exposure* to the ban. It is constructed as the share of comments in banned forums in the total number of comments left by a user in January-March 2020. α_i and τ_t are user- and week-fixed effects, respectively. $NumComments_{it}$ excludes comments left by treated users in banned forums (see Section 3 for details on variable construction). Thus, β reflects the average change in number of comments after the ban for a user who spent all their time in banned *subreddits* in January-March 2020. Throughout the analysis, standard errors are clustered at the user level.

Additionally, we estimate a version of Equation 1 with $NumberOfNewSubreddits_{it}$, referring to the number of new subreddits user i comments in in week t as the dependent variable. This measure provides a coarser, but easier for interpretation measure of real-location of activity across the platform. Coefficient on β in this case reflects the average change in number of new forums in which a user comments for a user who spent all their time in banned subreddits in January-March 2020.

Although our data allows for arbitrary time granularity, we opt for weekly aggregation and week-level fixed effect throughout. Analysis on weekly level allows us to abstract away from cyclicity in individual behaviour while still capturing short-term fluctuations in engagement due to the news cycle.

Identification In interpreting estimation results we follow the discussion in Callaway et al. (2024). We do not expect that "strong parallel trends" hold in this case. Specifically, while the policy change itself is exogenous, users still select into the pre-ban degree of treatment exposure: an increase in $ShareInBan_i$ reflects not only increase in treatment exposure, but also change in *the type of user*. We expect that different types of users likely have very different characteristics and thus potential responses to the policy. Therefore, the our estimates should be interpreted as the average treatment effect on the treated (ATT) on users who select into banned *subreddits*. This is, of course, precisely the effect of interest.

We split the continuous measure of individual policy exposure $ShareInBan_i$ into quartiles instead of following advice in Callaway et al. (2024) suggesting more parsimonious methods, such as various sieve estimators. The main reason is that we can clearly observe that parallel pre-trends do not hold for every exposure quartile. The results of sieve estimation would thus reflect effects of the policy for some groups and the pre-trend for others. Combining quartile split with an event study approach allows us to check whether the effect for a given subgroup is driven by the policy or the pre-trends. Although quartiles cannot capture user heterogeneity perfectly, we argue that they do reflect a meaningful difference in policy exposure.

We do not include zeros for the weeks in which a user has not been active on Reddit in the regressions, so the estimates can be thought of as intensive margin effects. Intuitively, we think that it is more appropriate to treat the decision on *whether to comment* on the platform in a given week as separate from the decision on *how much* to comment. From identification standpoint, the exclusion of zeroes leads us to ignore the effects due to platform exit; if the degree of user attrition from Reddit due to the policy change were high, our estimates could have been upward biased. However, as is shown in Section 4.2, we do not find evidence of substantial platform exit. Conservative Lee bounds analysis confirms that omission of zeroes cannot explain our results.

Sampling The equation is estimated both on the full sample and on the sample restricted to users who left at least 10 or 50 total comments on the platform between January and March 2020. The restriction applies symmetrically to both treated and control users. This exercise has two main purposes.

First is to ensure that the results are not driven by the artificially higher levels of policy exposure for users who only comment a handful of times overall and hence have small values of the denominator of $ShareInBan_i$. Figure A.2 displays the histogram of policy exposure by value of the denominator: there is clear bunching at specific values for users who left less than 10 comments overall.

Second is to address selection due to the definition of treatment group. Observing a treated user in any given week of the year *also implies* observing them at some point in weeks 1 to 14, since treatment definition is based on that period. Due to the natural lifecycle of accounts on social media, the pool of treated users is subject to constant attrition, and as the week of the year increases, the likelihood of observing a treated user who is also a more active user overall goes up. As a result, the effects of the policy on activity of treated users outside banned forums might be biased due to an upward trend in the data. Restricting the pool of users to those who left 10 or 50 comments in the first three months of the year helps mimic the selection pressure on the sample of treated users in the control group.

There is one particularity of estimation using restricted sample worth keeping in mind.

The cut-off on total number of comments left does not exclude activity in banned forums from denominator, but observations from banned forums are excluded from the regression data. That does mean that sampled treated users will mechanically have fewer observed comments than the control in weeks 1-14, and that difference is larger for higher quartiles of $ShareInBan_i$, since larger fraction of comments before the policy change is omitted. The pattern does not meaningfully affect the interpretation of results, but it does impact the pre-trends in the event studies for weeks 1-14, especially when logarithm of the number of comments is used as the outcome variable.

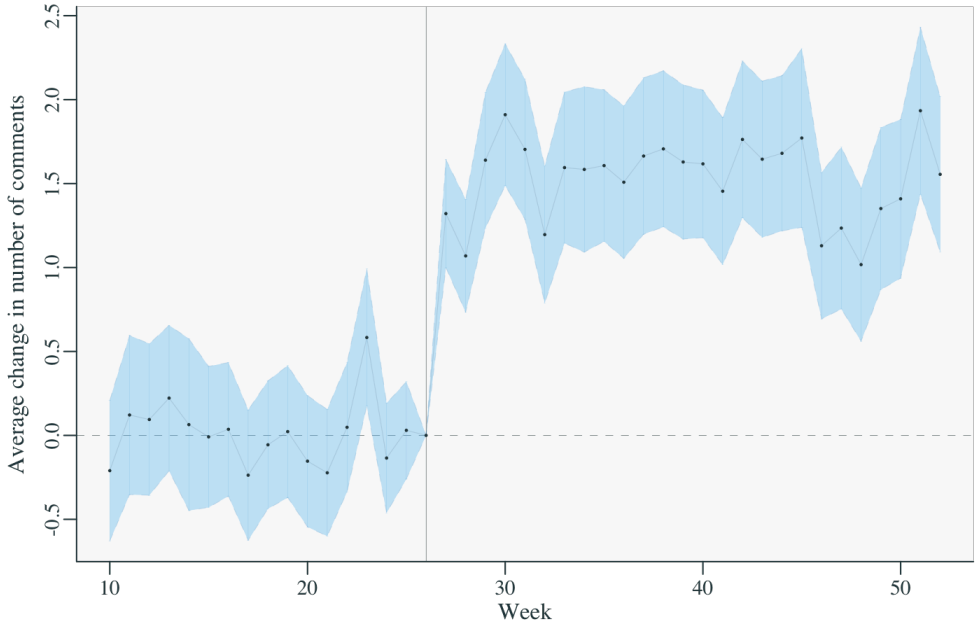
4.2 Main results

Displacement We find that the ban leads to an immediate, significant increase in the number of comments left by *core* treated users (users in the top quartile of share of comments left in banned forums) on Reddit. Figure 1 presents the event study of the effect on the number of comments left in each week, showing that the policy change led to an average increase in activity by approximately two comments. Estimation using logarithm of the number of comments as the outcome variable (Figure B.15) suggests that the effect corresponds to an approximate 10% average increase in number comments left by a treated user outside banned forums. Figure B.1 shows the effect in the raw data.

The increase in activity is sustained in the long run, suggesting that it is driven by *substitution* to other forums, rather then short-term backlash to the policy change. A temporary increase in number of comments visible in both event studies and in the raw data three weeks prior to the policy change corresponds to the week after the killing of George Floyd. The reaction is clearly short-term and small relative to the effects of the policy change itself, alleviating the concern that the effects we observe are attributable to the preceding political movement itself and not to the moderation intervention that it precipitated. It also provides an intuitive benchmark for the significance of the effect within the context of Reddit– the policy change leads to a persistent increase in activity of *core* users relative to the other users of the platform that is several times larger than that of the previous news shock.

Estimates for other quartiles are shown in Figure B.16. As was discussed in the previous section, we expect that pre-trends do *not* hold for each subsample of the users due to the selection pressure on the treatment group. This is indeed what we observe: in all specifications the event study estimates for the first quartile are dominated by a large downward linear trend, and parallel trends for the second and third quartile only hold after we impose restriction on the total number of comments left in the first 3 months of the year. Still, the variety of estimations paint the same general picture: robust and persistent increase in activity on the rest of the platform for the *core* users, slight short-term increase for the third quartile and slight negative coefficient on the second quartile, likely reflecting the pre-trend. Overall, we see that the ban did not have a substantial

Figure 1: Effect of the policy change on the number of comments left by *core* treated users: difference-in-differences estimates



Notes: Figure 1 shows the event study of the effect of the policy change on the number of comments left by *core* treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level, N=232,126,419. Equivalent results using log of number of comments as the outcome variable are presented in Figure B.15.

effect on the treated users whose normal Reddit use was "disrupted" by the policy change to a relatively lesser extent. We argue that the correspondence between higher exposure to the policy change and larger increase in comments on Reddit after the policy change is indicative of the substitution as the channel for the effects that we observe.

Migration To establish whether substitution is driven by increased activity in previously visited forums or increased activity in new *subreddits*, we estimate a specification equivalent to that of Equation 1, but on the panel of the number of comments a user left in a week in a specific *forum* with user-by-forum instead of user-fixed effects (see Equation 2 below).

$$NumComments_{ist} = \alpha_{is} + \tau_t + \beta Post_t \times ShareInBan_i + \epsilon_{ist} \tag{2}$$

The estimation allows us to look at the average change relative to pre-policy activity level within the same forum set. Additionally, we estimate Equation 1 using number of new forums in which user was active per week as the outcome variable.

Figure B.6 shows the results for the activity in previously visited forums, Figure B.5 displays event study estimates for the number of new forums. Once again, we focus on *core* users for which we observe a an increase in overall activity after the ban. The policy

change leads to both a persistent increase in comments left in previously visited forums and an increase in activity in new *subreddits* in the first ten weeks after the policy change. The results suggest that the ban of hateful forums leads to a substantial *disruption* in the way treated users interact with the platform. The ban of hateful forums induces *migration* across the platform resulting not only in increase in activity in previously used forums, but also a search of substitutes. Thus, it is likely that the shock to "consumption basket" of treated users likely had additional ripple effects on other users of Reddit, whose level of exposure to migrating treated users changed after the ban.

Since we do not have clickstream data and rely on comments to measure user activity in a particular location, our results do not reflect the full extent of content consumption on Reddit, only its *production*. In that sense, the full extent of reallocation is likely larger in magnitude. Additionally, the results on the number of new forums in which a user comments in a week reflect two separate behavioural dynamics. One is "successful search" of *subreddits*-substitutes, whereupon a user chooses to comment in a forum they have not visited previously. The other is the decision to go from passive consumption of forum content which user read prior to the policy change (or, as it is usually referred to, *lurking*) to leaving comments. We cannot disentangle the two, but can interpret them jointly as an increase in number of forums in which user participates actively.

Platform exit An important question is whether the policy leads some users to leave Reddit altogether. Previous literature has reported that moderation interventions can induce most toxic users to migrate to less regulated places, such as in the case of migration of right-wing users from Twitter/X to Parler, the largest right-wing alternative platform at the time, documented by Rizzi (2024). Aliapoulios et al. 2021 analysing all shocks to Parler's membership do report an influx in users in response to increased Twitter moderation during the George Floyd protests, but do not mention a significant increase related to the ban of hateful forums on Reddit.

Another relevant discussion is the degree to which moderation intervention aligns with platform's interests. Reddit, as all other major platforms, is monetized primarily through ad shows. A moderation policy that triggers substantial backlash and disengagement can have a negative effect on the social media platform's bottom line.

A disproportionately large increase in drop-off of treated users from Reddit might also signal that the increase in number of comments left on Reddit by *core* users after the policy change that we observe is driven by selection of more active users into the sample, not within-user substitution.

Figure B.2 shows the fraction of all treated and untreated users last observed in each week of 2020. To mimic the selection criteria for the treated group, we restrict the sample of control users to those present between January and March 2020 as well. Surprisingly, there is no observable difference in average levels of drop-off from the platform between

the treated and control groups in the weeks around policy change. In Figure B.3 we zoom in on the period of 20 weeks around the policy change and split the sample of treated users into quartiles of policy exposure. There is a slight increase in attrition of *core* treated users in the first week after the ban, but the difference is very small, around 0.5% absolute increase or a third of the average "normal" levels of weekly platform attrition for treated users. Still, we compute conservative Lee bounds on the estimates of activity increase (Lee 2009) to ascertain that platform exit does not drive our results (Figure B.30).

Why do not we observe platform exit as some work on other platforms does? In general, we argue that the degree of user attrition is very much dependent on the context of the moderation intervention, especially on the specific *costs of off-platform substitution* that users affected by a moderation intervention face. The majority of available less regulated alternatives at the time of the policy change were targeted towards extreme right-wing users (4chan, Parler, Gab, Voat and others). For at least half of the users in the treated sample none of these platforms are a natural substitute to Reddit.

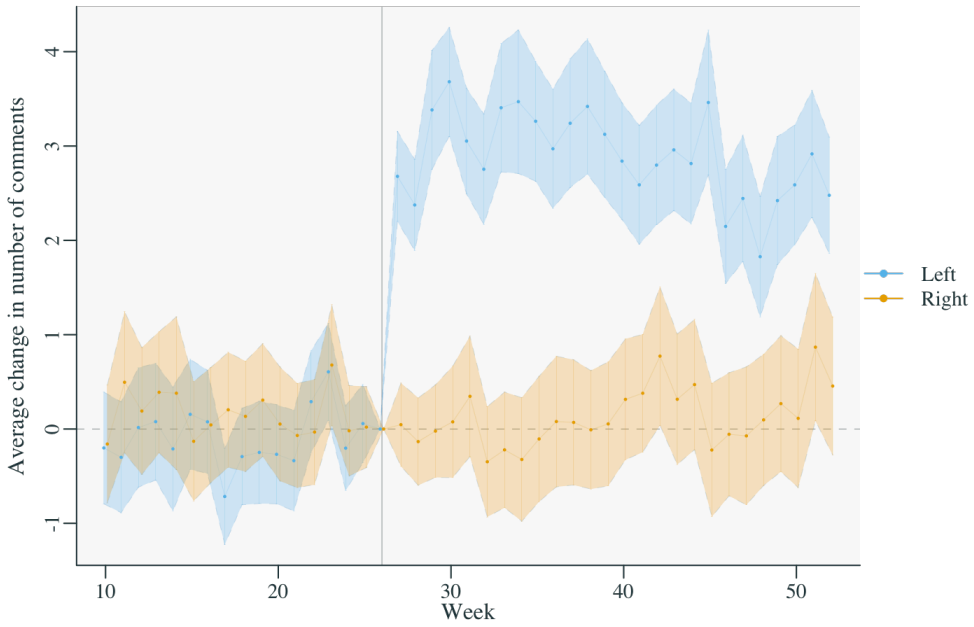
Additionally, the vast majority of moderation policies for which platform exit can be analysed in principle do not imply a well-defined and easily identifiable set of affected users. Hence, observed platform exit reflects *selection into exit as policy response* that is not uniform, but correlated with individual substitution costs. Our results, on the other hand, pertain to attrition as a fraction of a well-defined treated user set. We do not argue that platform exit does not happen, just that it does not constitute a meaningfully large part of policy response. We discuss heterogeneity in platform attrition and the likely components of substitution costs for different user groups in more detail in subsequent sections.

Robustness Figure B.17 and B.18 present estimates of the main results based on the sample restricted to users who left at least 10 or 50 comments on the platform in January-March 2020. The estimates confirm our previous finding that the policy change leads to an increase in activity of *core* users and a short-term increase for users in the third quartile. Robustness to alternative sampling procedure and clear sharp increase in the coefficient estimates in the week after the policy suggest that the overall increase is not driven by selection pressure in the sample of treated users (see Section 4.1 for discussion). It also suggests that the effects are not driven by oversampling users with low baseline level of activity in the upper quartiles of policy exposure ($ShareInBan_i$).

4.3 Heterogeneity

We now present results for different groups of treated users depending on the type of banned forum they were active in. As described in Section 3, we do this by constructing indicators of primary topic for the banned forums and then assigning users in the corresponding category if more than 90% of their comments in banned forums are left in forums of the

Figure 2: Effect of the policy change on the number of comments left by *core* left- and right-wing treated users: difference-in-differences estimates



Notes: Figure 2 shows the event study of the effect of the policy change on the number of comments left by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using log of number of comments as the outcome variable are presented in Figure B.23. Results from a triple difference specification are shown in Figure B.24.

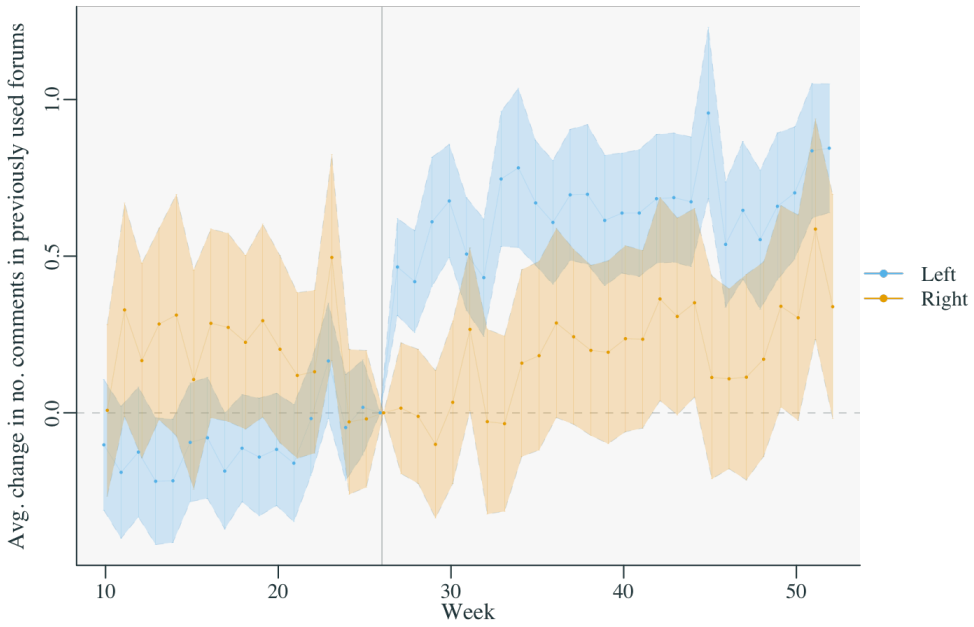
corresponding type. Since most *subreddits* can be designated as politically right- or left-wing, we first consider this grouping.

Figure 2 shows the results of re-estimating Equation 1 separately for left- and right-wing treated users. Since the increase in activity after the policy change is driven by the *core* users, we focus the attention of this group.

We find a striking difference in response to the ban depending on the type of banned forum users were active in: the overall effect is driven entirely by the *core* left-leaning users who leave, on average, 2-3 more comments on Reddit after the policy change – approximately 15% increase relative to control group (Figure B.23 shows the results with logarithm of the outcome variable). There is no corresponding change in activity of *core* right-wing users.

The results of estimating Equation 1 using number of new forums in which a user commented each week as outcome variable (Figure 4) are symmetric: the increase that we find in aggregate results is largely attributable to the left-wing users, with a small increase for *core* right-wing users in the week after the policy likely attributable to discussions *about* the ban in new forums. The same relationship holds for the effect on number of comments left within previously used forums (Figure B.25).

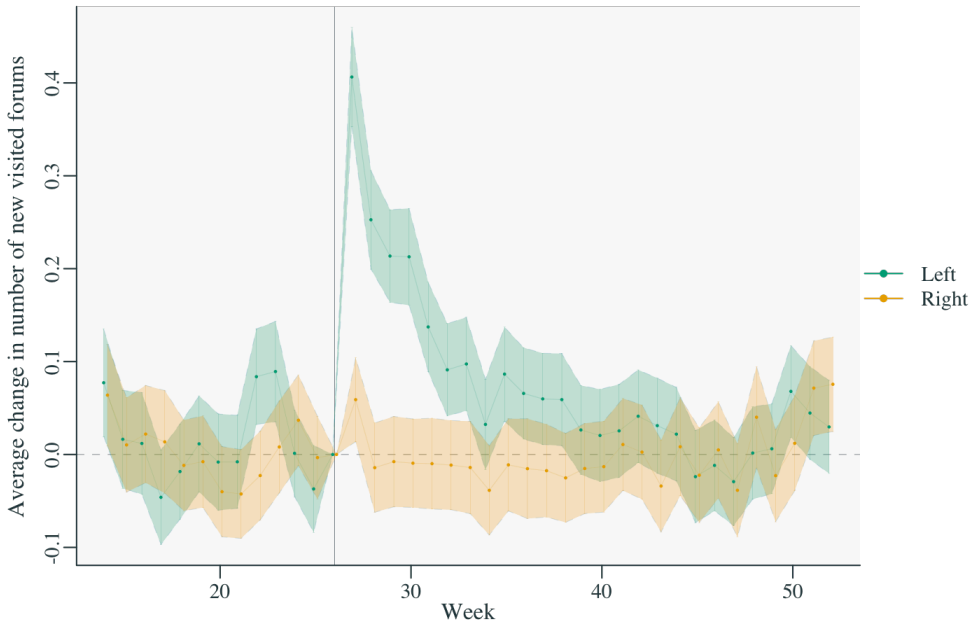
Figure 3: Effect of the policy change on the number of comments left in previously used forums by *core* left- and right-wing treated users: difference-in-differences estimates



Notes: Figure 3 shows the event study of the effect of the policy change on the number of comments left in previously used forums by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the average change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

The magnitude of the difference between the two groups is surprising: cutoffs on activity in banned forums determining *core* users are the same for both groups and the results are robust to estimation using logarithm of number of comments as outcome variable (Figure B.23), as well as restricting the sample to those who comment at least 10 or 50 times in January-March 2020 (Figure B.21 and Figure B.22, respectively). This implies that left- and right-wing users who experience a shock to normal Reddit use of a similar magnitude do not react to it in the same way.

Figure 4: Effect of the policy change on the number of new forums visited by *core* left- and right-wing treated users: difference-in-differences estimates



Notes: Figure 4 shows the event study of the effect of the policy change on the number of new forums visited by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4$^{\text{th}}$ quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the average change in number of new forums visited by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Our interpretation is that left- and right-wing users face drastically different *substitution costs* within and outside Reddit. The exact nature of these substitution costs might vary: perhaps, there are no suitable alternatives to the right-wing banned forums on Reddit or the search of these substitutes is prohibitively costly. Given that we observe a null effect on activity in new forums and on the number of comments left in previously visited *subreddits*, it appears likely that the distance between the types of discussions in right-wing banned forums and the rest of the platform is larger than that of the left-wing ones.

Lack of substitution for right-wing users within Reddit can be accompanied by substitution of discussions from banned forums to alternative platforms. At the same time, we do not see evidence of disproportional attrition of right-wing users from Reddit or decrease in the overall activity outside banned forums in the long-run, suggesting that even if off-platform substitution happens in the first place, it does not negatively impact engagement of *core* right-wing treated users on Reddit beyond hateful forums.

Reddit is a general-interest platform and treated users are engaged in a variety of communities, often completely different from one another. In fact, distribution of activity of treated users across non-banned forums is very similar to that of other users of Reddit. It is plausible that lack of reaction to the policy change reflects that the value of Reddit as a platform for right-wing users is not completely determined by availability of hateful forums. Alternative platforms can provide a space for discussions similar to that of banned

forums, but it is unlikely that they can replace the general interest aspect of Reddit use.

Right-wing banned forums are a lot more heterogeneous than the left in terms of their topic, encompassing alt-right, trans-exclusionary, generally offensive and other types of content. For that reason, an alternative viable explanation is that, similarly to the dynamic between overall effects and effects disaggregated by left- and right-wing users, the estimates for the right-wing treated users are masking internal heterogeneity in response to the policy change between subgroups of treated right-wing users.

We now consider a more detailed grouping of right-wing treated users, additionally splitting the sample into users of transphobic, alt-right, generally offensive forums, as well as users of `r/The_Donald` (see Section 3 for details).

Figure B.26 presents the results of estimating Equation 1 separately for each subgroup; results using log number of comments as outcome variable are presented in Figure B.27. As before, we are zooming in on the *core* users of banned forums. The first row presents results for right- and left-leaning users discussed in the previous section as a reference point, the latter two rows show estimates by the four subgroups of right-wing treated users.

While the pre-trends do not hold for every subgroup in the baseline specification, the general picture is clear. The null result observed for right-wing group overall holds for users of subreddits containing offensive humour and users of alt-right forums excluding `r/The_Donald`.[7] However, there is a clear increase in number of comments left by users of transphobic forums. The confidence intervals are wider, but the magnitude is comparable to that of left-wing users and the effect is similarly sustained until the end of the year.

`r/The_Donald` The results for the users of `r/The_Donald`, by far the most notorious banned forum, present an interesting case study in itself. We see a slight decrease in activity on Reddit after the policy change. The decrease is more pronounced in specifications with cut-offs on total activity in January-March 2020, suggesting that they are driven by more active users of the platform. But most importantly, we observe a clear jump in activity between weeks 8 and 10, again similar in magnitude to that of the jump for left-wing users after the ban and persistent in the long run. Similar results emerge from specification with number of new forums visited as outcome variable shown in Figure B.28: we see evidence of activity in new forums similar in magnitude and duration to that of *core* left-wing users – but happening in weeks 8-10, more than four months prior to the policy change.

In fact, the jump that we observe coincides with a ban of a number of `r/The_Donald` moderators and placement of the subreddit in "restricted mode" which limited the ability of most users to leave new posts on the *subreddit* on February 26, 2020 (week 9 of our sample). Users of `r/The_Donald` are affected by a different moderation intervention restricting their ability to use the forum earlier than the rest of the sample. Is not surprising that we do

[7] We also observe that, as expected, the spike in activity in the weeks after George Floyd's murder for right-wing users is entirely due to users of r/The_Donald.

not see a significant effect of the policy change on users of r/The_Donald in June 2020 – their substitution to other forums on Reddit already happened at that point.

`r/The_Donald` has attracted significant media attention due to it's size, degree of toxicity, and role in the online political discourse of the US in 2010s. Many outlets claimed that the June 2020 ban was meaningless, since users of the forum have created an alternative platform (`thedonald.win`) with the sole purpose of replacing the original Reddit community quarantined in June 2019.

Simple descriptive statistics suggest that `r/The_Donald` was far from inactive in the beginning of 2020: with almost 1.5 million comments left in January-March 2020, it is the second largest forum in our sample (Figure A.1).[8] Still, we might expect that, given the existence of a well-known alternative outside Reddit, restriction of the forums in February 2020 would have little effect on migration of its users within Reddit.

Judging by the event study estimates, this is clearly not the case. The pattern is complementary to findings in Horta Ribeiro et al. (2021) who report that users migrating to the alternative website after subreddit was restricted were on average less active and more toxic on the alternative website. This result is important in the context of the discussion about adverse effects of moderation pushing users into less regulated platforms. r/The_Donald and its successor thedonald.win are *the main example* of such movement in the literature on the topic. And still, the dynamic is not universal: our results show that migration can be less pronounced for users who value platform itself to a relatively higher extent.

4.4 Mechanisms

To summarize, we find that most active treated users increase their activity in other *subreddits* after the ban. As is clear from the event studies, the effect is sustained in the long run, suggesting the results cannot be explained by the initial reaction to the policy change itself. Our interpretation is that the ban leads to substitution of banned forums to the rest of the platform; naturally, the magnitude of substitution for the *core* users of banned forums most affected by the policy change is the largest. At the same time, substitution patterns appear to vary by user group: the effect we see is driven entirely by left-wing and transphobic users. But where do treated users actually go?

Comovement Increase in activity of *core* users could be associated with re-creation of banned communities elsewhere on Reddit. If such new communities are similar to banned forums in the degree to which they promote hateful content, that would imply that the policy was ineffective at dealing with hate speech. Furthermore, if whole groups of treated

[8] Chandrasekharan et al. (2022) analyse the effect of quarantine of r/The_Donald in detail. Although the estimates in the paper are likely affected by identification issues, the results suggest that quarantine had little effect on activity of existing users or use of hate speech, but limited the influx of new users. The authors find scarce evidence that moderators of the forum encouraged migration before the ban.

users migrate into existing *subreddits*, the ban could result in negative spillover effects on existing users of "invaded" forums or even their displacement.

To test whether networks of treated users are re-created elsewhere on Reddit, we estimate the effects of the policy on the probability that treated users are interacting with each other outside banned forums. If banned communities are re-created somewhere on the platform, we should see an increase in probability that a treated user is replying to another treated user relative to probability that a non-treated user replies to a treated one.

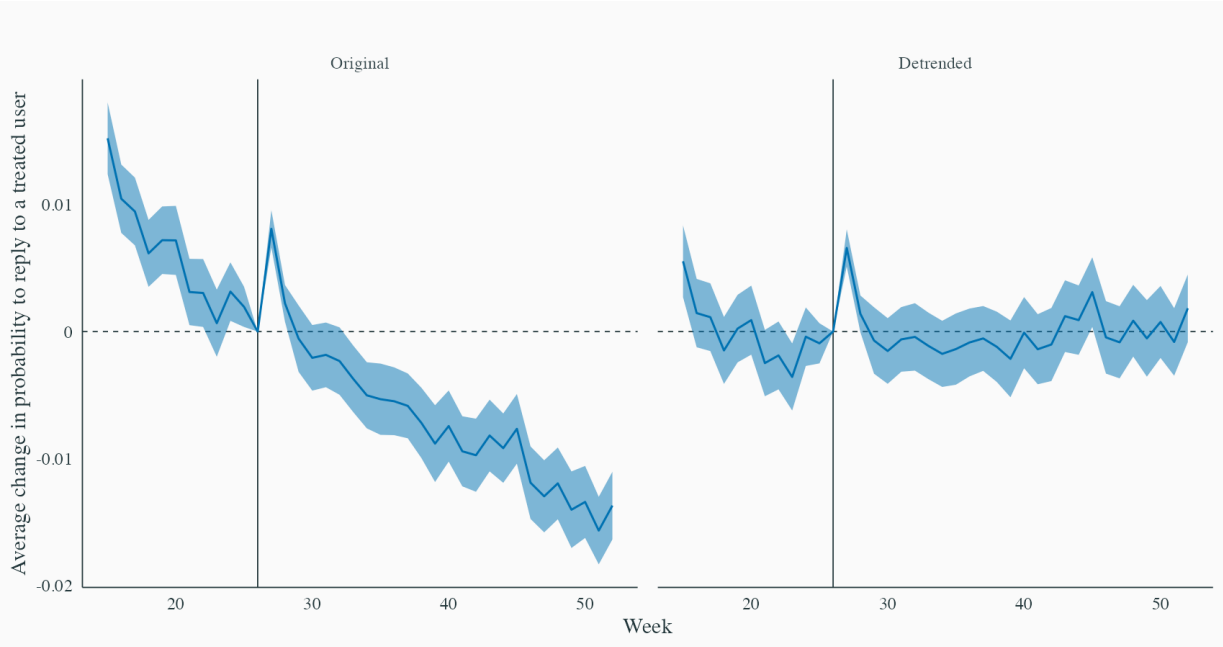
We estimate a comment-level difference-in-differences specification (Equation 3) with user-by-subreddit and week fixed effects. We restrict the sample to users who left at least 50 comments in the first 3 months of the year (it is computationally impossible for us to produce event study estimates on the whole sample). The outcome variable is a dummy equal to one if a comment is left as a reply to a treated user. β thus reflects an average change in probability that a treated user is interacting with another treated user outside banned forums.

$$IsReplyTreated_{cist} = \alpha_{is} + \tau_t + \beta Post_t \times Treated_i + \epsilon_c \tag{3}$$

As before, $Post_t$ is a dummy variable equal to one in the weeks after the ban and zero otherwise, $Treated_i$ is 1 for treated users and 0 otherwise.

Figure 5 shows the results. Panel A displays original event study coefficients, Panel B shows the results after partialling out the linear trend. Raw data is presented in Figure B.4. The linear trend in this case is a feature of the data reflecting a mechanical decrease in probability of interacting with a treated user as the total pool of treated users shrinks over the weeks of the year. All three figures show the same pattern: a sharp increase in the probability that two treated users are interacting outside banned forums in the week after the policy change and no effect afterwards. The initial "spike" likely reflects users talking *about the ban* in forums such as *r/subredditdrama*.

Figure 5: Effect of the policy change on the probability that a comment is reply to a treated user



Notes: Figure 5 presents the estimates of the effect on the policy on the probability that a given comment of a treated user is a reply to a comment left by another treated user. The results are obtained by estimating comment-level Equation 3 with a dummy for whether the comment is a reply to a treated user as the dependent variable. The sample is restricted to users who posted at least 50 times between January and March 2020. Left panel presents raw estimates, right panel presents estimates after partialling out the linear trend. The estimates show a clear temporary increase in probability that two treated users are interacting outside banned forums after the policy change. Confidence intervals are at 95% coverage, errors are clustered at the user level, N=384,020,091.

To alleviate the concern that estimates from an event study specification fail to detect an increase in comovement due to the small size of the treated group relative to control, we also present detailed raw data plots on the fraction of comments of treated users left as reply to each relevant user group. The proportion of total comments that are interactions among treated users stays almost constant for the left-wing users throughout the year and gradually decreases after the policy change for the right-wing subsample. It is worth noting that the share of interactions with other treated users outside banned forums is actually quite small at around 5%, reflecting the fact that for most treated users normal use of Reddit encompasses much more than just interactions in banned forums.

Overall, across exercises we do not see a sustained increase in the probability that two treated users interact outside banned forums, therefore we do not find evidence that networks of treated users are recreated outside banned forums after the policy. This does not rule out the possibility that some treated users find each other after the policy change and continue interacting outside banned forums, only that we do not observe coordinated re-creation of banned communities elsewhere on Reddit on the scale that would put into question the overall impact of the policy change.

This also suggests that the potential for large-scale spillovers from the policy change

is limited. The overall significance of spillover effects depends on the magnitude of effect of interacting with a treated user on an individual level and the volume of additional interactions with treated users due to the policy. We cannot say anything about the former (at this point), but the *possible reach* of the policy in terms of exposing users to migrants from hateful forums is limited. Firstly, raw data shows that the probability that a non-treated user replies to a treated user does not increase after the policy. Secondly, the very act of breaking up the networks of users in banned *subreddits* over time forces a substantial drop in the probability of encountering these users: no new treated users are "created", and the normal gradual attrition from the platform constantly decreases their total number.

Drivers of migration So far, we have established that substitution of banned forums is not associated with re-creation of their user networks elsewhere on the platform. But substitution can have other contributing factors as well, such as "closeness" of substitutes to the banned forums.

All banned forums exhibit high level of toxicity and hate speech use, most of them are centred on politically sensitive topics. In view of the argument in Beknazar-Yuzbashev et al. (2025) on the tight link between toxicity and engagement, it is informative to try to understand whether treated users are interested in specific discussion topics which tend to be toxic by their nature or seek out generally emotionally engaging conversations, which tend to happen in politically sensitive areas, and exhibit high degree of toxicity and hate speech use.

We cannot measure inherent propensity of content to elicit engagement, but in this setting we can try to approach the problem from the other end. Here, we consider previous individual level of attention received from *other users* of a forum as a measure of engagement.

We construct a panel of the number of comments left by each user in a given forum in a given week, as well as a suite of forum characteristics capturing levels of toxicity and user engagement. We then estimate the following difference-in-differences specification:

$$NumComments_{ist} = \alpha_{is} + \tau_t + \beta Post_t \times ShareInBan_i \times \Gamma_{(i)s} + \epsilon_{ist} \tag{4}$$

Where $\Gamma_{(i)s}$ is a placeholder variable referring to a set of *subreddit* (Γ_s) or *subreddit*-by-user (Γ_{is}) characteristics. $NumComments_{ist}$ is the (log) number of comments left by user i in subreddit s in week t. Treatment variable $Post_t$ and $ShareInBan_i$ are defined as before.

We cannot interpret results of this estimation causally: users increasing activity in most toxic forums do not necessarily do that *because* the forums are toxic. But we do expect users to comment disproportionately more in forums containing hate speech if direction of substitution is related to the similarity of discussion topics to those of banned forums.

We find clear evidence that *core* treated left-wing users increase activity in more toxic forums and forums with higher incidence of hate speech from the set of forums they have already visited before the policy change (Figure B.8 and Figure B.10 present the results).

At the same time, we do not find strong evidence that increase in activity falls disproportionately on forums where a treated user can elicit engagement from other users. Figure B.12 shows the effects on activity by average score of comments of a treated user in a forum, Figures B.13 and B.14 show the effects by number of replies elicited by a user in a forum prior to the policy change.

Taken together our evidence does not support the idea that substitution is driven by level of engagement that treated users receive from others, more fundamental similarity to banned forums appears to be correlated to post-policy change activity increase to a higher extent.

5 Effects on use of hate speech

5.1 Empirical strategy

In the second half of the paper we are interested in the effects on use of hate speech – primary target of the moderation policy. As was discussed in Section 3, we use incidence of slurs as proxy for hate speech use. The main comment-level specification is presented in Equation 5.

$$NumSlurs_{cijst} = \alpha_{i(s)} + \tau_t + \gamma_j + \beta Post_t \times ShareInBan_i + \epsilon_{cijst} \tag{5}$$

Here, $NumSlurs_{cijst}$ stands for number of *unique* slurs in comment c by user i under post j in subreddit s and week t. We only count the first use of a slur in any given comment to adjust for likely autocorrelation between instances of the same slur within a comment.[9] As before, $Post_t$ is a dummy variable equal to one in the weeks after the ban and zero otherwise, $ShareInBan_i$ is the share of activity in the banned subreddits in total user's activity in January to March 2020. τ_t is week-fixed effect, $\alpha_{i(s)}$ is a user- or user-subreddit fixed effect, depending on specification.

Additionally we estimate the main equation with post- or thread-fixed effects γ_j: that is, we control for the baseline level of hate speech use in the discussion where the comment was left. It is computationally impossible for us to include both user-subreddit α_{is} and post or thread γ_j fixed effects in the same specification due to the size of the sample. However, all specifications separately produce very similar results.

We do not take logs of number of slurs, since there are many zeroes in the data. As suggested by Chen and Roth (2024), we instead look at the absolute changes and normalize them by the mean of the dependent variable for treated users before the policy change to obtain effect magnitudes. Since the incidence of slurs is low relative to the number of comments, we multiply coefficients by 1000 to obtain effects in "slurs per 1000 comments" for ease of presentation and interpretation; the transformation does not affect the results. We also estimate the effects using Poisson QML estimator (see notes to Table C.5 for the discussion on the problems arising when using Poisson for data with large number of groups containing only zeroes).

5.2 Main results

Reduction in use of slurs Table 1 shows the results of estimating Equation 5 with incidence of slurs as the outcome variable on the full sample of comments. Columns (1) to (4) include progressively more granular fixed effects to account for selection into conversations on different topics and into different communication norms.

[9] An intuitive way of thinking about this is to distinguish the decision to use a slur in a comment from the decision on how many instances of a slur to use conditional on having decided to use it.

Table 1: Effects on incidence of hate speech: difference-in-differences estimates

	Number of unique slurs per 1000 comments			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i$	-0.388***	-0.307***	-0.395***	-0.324*
	(0.057)	(0.106)	(0.073)	(0.129)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	28,979,600			
User-subreddit pairs		242,242,341		
Posts			158,682,911	
Threads				908,670,654
N	1,884,546,944	1,884,546,944	1,884,546,944	1,884,546,944
Dep. var. mean	1.82	1.82	1.82	1.82
R^2	0.05	0.18	0.13	0.52

Notes: Table 1 presents the results of estimating Equation 5 across all quartiles by policy exposure. For ease of display the dependent variable is multiplied by 1000. We only count unique words due to probable autocorrelation in use of specific language within a comment. The effects by quartile of policy exposure are presented in Table C.2. Event study based on specification in Column (1) is presented in Figure C.7. The results from Poisson QML estimation are presented in Table C.5. The stability of the coefficients in specification with ultra-granular post- and thread-fixed effects suggests that the results are not explained by treated users sorting into discussions with lower average levels of hate speech use. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

We document that the ban reduces the incidence of slurs in the comments of treated users by 17 to 21% relative to platform mean. It is important to stress that the effect is *not a mechanical consequence* of the ban of toxic forums, because observations from banned *subreddits* are completely excluded from the analysis; the effects reflect the reduction in hate speech use *on the rest of the platform*. Results of estimation with an indicator for whether a comment contains slurs as dependent variable presented in Table C.4 are similar but slightly larger in magnitude (21.6 to 24.9% relative to the mean), suggesting that the effect is largely on the extensive margin of slur use. The same pattern emerges in an event study using specification identical to Column (2) of Table 1 on the computationally feasible subsample of users who left at least 50 comments between January and March 2020 (Figure C.7), as well as in the raw data (Figure C.1).

Since incidence of slurs represents count data, we additionally validate that the results

are robust in Poisson quasi-maximum likelihood estimation. The results presented in Table C.5 are robust but are smaller in magnitude. The reason is that Poisson QML estimator implicitly drops groups which contain only zeroes, leading to a large increase in the estimate of the average incidence of hate speech and smaller absolute magnitudes of the coefficients estimates.

Selection into toxic topics A prominent feature emerging from the data is the degree to which the use of toxic language and hate speech is *space-dependent*. Degree of toxicity and hate speech use varies significantly across and even within *subreddits*, and users adjust their language significantly to the norms of the forum in which they comment. This is reflected in the statistics describing the use of slurs by treated users in an outside banned forums: for most slur categories, incidence of slur use within banned forums is at least double that of non-banned, for the same underlying sample of users (Table A.6). In fact, as we discuss in more detail in sections on spillover effects (Section 5.4) and mechanisms (Section 5.5), indicator for whether a comment is a reply to slur is a strong predictor of hate speech incidence (see Table C.19 for estimates).

We explain this concentration in hate speech use by a combination of what we call *selection and contagion*. Selection means that similar users self-select into participating in similar conversations, and high correlation in toxicity of replies in the same discussion reflects partly that we are more likely to see two toxic users together. Contagion, on the other hand, is the higher likelihood to reply with a slur after being exposed to a slur *conditional on user type*.

Hence, the overall reduction in use of slurs that we observe can be driven not only by change in behaviour within the same types of discussions, but also by *sorting* into forums or conversations with lower average use of hate speech. Selection does not have to be *on* hate speech incidence directly, selection on topics which happen to have a lower average "intrinsic" hate speech incidence would produce the same result.

Given the results on overall volume of activity of displaced users, suggesting that at least a subpopulation of users migrates to other forums in the aftermath of the ban, it is plausible that ban of hateful forums can affect the types of discussions treated users choose to participate. For example, reduction of exposure to banned forums can lower the salience of the corresponding topics and attention to them outside banned forums, especially if the network of treated users in banned forums facilitates search of toxic conversations across the platform. Conversely, we could observe higher selection into toxicity and hate speech if they are correlated with topics of forums-substitutes.

The near-population Reddit data allows us to disentangle the two channels by incorporating ultra-granular fixed effects. Column 2 of Table 1 includes a fixed effect for each user-subreddit pair, Columns 3 and 4 include post- and thread-fixed effects, respectively. We can thus estimate the effect on use of slurs *controlling for local level of hate speech use*.

The magnitude of the coefficients decreases only slightly in specifications with post- and thread-level fixed effects: it is clear that the effect is largely driven by lower incidence of hate speech use relative to control users in the same discussion types, not by sorting into discussions involving less hate speech.[10]

Still, it might be the case that the policy change influences selection into conversation types in a way that is not captured by the level of hate speech use. Since we do not know which particular dimensions might affect selection, we cannot measure them directly. However, if we are willing to assume that observing a treated user in conversation with the same control user means observing them in a similar conversation type, we can try to correct for unobserved selection by focusing on the effects partialling out level of hate speech use of a *user pair*. Table C.22 estimates Equation 5 with user pair fixed effects. The estimates are robust with slightly larger magnitudes than in the main specification, indicating that the decrease in hate speech use after the policy change persists *in interactions of the same two people*. Overall, stability of the coefficients across specifications suggests that the results we observe are not driven by selection but rather reflect *within-user language change*.

Interactions of treated users Table C.21 estimates the effects on the use of slurs additionally interacting the independent variable with a dummy for whether comment is a reply to a treated user. The results show that, even after adjusting for selection into different baseline use of slurs by user in a subreddit (Column 2), replying to another treated user has an additional negative effect on the incidence of slurs after the policy change. Inclusion of pair-fixed effects allows us to additionally see whether the effect persists in interactions with the *same users* before an after the policy change. While the main coefficient in this specification remains stable - that is, treated users use less hate speech after the ban even when talking *to the same people*, the interaction with dummy for reply to a treated user does not have an additional effect, likely due to low power or if pair-fixed effect captures contagion near perfectly.

The estimates alleviate the potential concern that the overall decrease in hate speech use we observe is associated with an increase in hate speech use in interactions of treated users among themselves. Instead, we observe a significant additional decrease in hate speech use when a treated user replies to another treated relative to control.

Effects on *core* users We have seen in the previous section that degree of substitution in the aftermath of the ban is higher for users with large magnitude of individual exposure to the policy change. Do we observe the same pattern for language change? Table C.2 disaggregates the results shown in Columns (1) and (2) of Table 1 by quartile of user activity in banned forums. To ensure that the results are not driven by users who leave few comments before the policy change, we estimate the same specification after restricting

[10]The increase in R^2 from specification with post to thread fixed effects further reinforces the idea that use of hate speech on Reddit is very concentrated in specific discussion threads.

the sample to users who left at least 50 comments in January-March 2020 and obtain similar results.

Although the coefficient estimates are negative in all cases, we do not observe systematic correlation between degree of policy exposure and the magnitude of the reduction in hate speech use. Coefficients on users in the first quartile are relatively smaller and are not significant after controlling for hate speech use of a forum in which they are active. Interestingly, the largest and most robust relative effects we observe are on the second and third quartile of activity in banned forums. The magnitudes for *core* users of banned forums are much lower in absolute and relative sense and are not significant in all specifications. It is clear that while migration across the platform is driven by degree of activity in banned forums, the effect on use of hate speech does not follow the same pattern.

Given descriptive statistics, this is not entirely surprising: Figure C.4 and Table A.8 both show that, although the correlation between degree of policy exposure and slur use is positive, it is small in magnitude, especially for left-wing users. In other words, degree of pre-ban activity in banned forums determines policy exposure, but does not necessarily reflect a specific user type in terms of preferences for hate speech use. This observation agrees with our previous suggestion that within Reddit specific conversation space matters more than characteristics of individual users. We discuss the likely explanations for this pattern in Section 5.5.

Another relevant consideration here is the distinction between moderation policy exposure in terms of production and consumption of content of banned forums. $ShareInBan_i$ reflects the degree to which the ban shocks the normal volume of comments left by individual user but it does not say much about how much content of banned forums a user was consuming passively.

5.3 Heterogeneity

In this section, we analyse heterogeneity of the estimates by user groups and slur types to provide suggestive evidence on the drivers of the overall reduction in hate speech use.

Table 2 shows the results of estimating Equation 5 after interacting independent variable with dummy for polarity of the treated users. Unlike the effects on activity on the platform, the effect on hate speech use is largely driven by right-wing users, and the coefficients on the left are much smaller in magnitude and largely insignificant, even after accounting for the fact that the baseline use of slurs by left-wing users is lower than for the right-wing ones.

In Table C.10 we estimate the specification in Column (2) separately by category of slur. Similarly to the main estimation results, the group-level estimates for left-wing users suggest no effect on use of hate speech. As for the right-wing users, we observe that the reductions occur in all slur groups, but not in a uniform way: the largest *relative* effects are on antisemitic (80% of platform mean), homophobic (69%) and transphobic (102.5%)

language.

Remarkably, comparison with descriptive statistics (Table A.7 shows that these are *not* the categories for which average slur use by treated right-wing users outside banned forums was the highest, but the ones for which it was *the lowest*. In fact, with the exception of racial and ethnic slurs, *the larger the difference between slur use between banned and non-banned forums for a slur type, the larger the magnitude of the effect that we observe.*

Table 2: Effects on incidence of hate speech by political leaning of the users: difference-in-differences estimates

	Number of unique slurs per 1000 comments			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i \times Left_i$	-0.087*	-0.047	0.064	0.042
	(0.040)	(0.045)	(0.050)	(0.062)
$Post_t \times Treated_i \times Right_i$	-0.393***	-0.319*	-0.425***	-0.266
	(0.075)	(0.145)	(0.092)	(0.170)
$Treated_i \times Left_i$			-0.144	-0.127
			(0.126)	(0.128)
$Treated_i \times Right_i$			0.994***	0.801***
			(0.141)	(0.160)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	28,963,687			
User-subreddit pairs		240,919,975		
Posts			158,244,079	
Threads				905,753,130
N	1,875,279,908	1,875,279,908	1,875,279,908	1,875,279,908
Dep. var. mean	1.82	1.82	1.82	1.82
R^2	0.05	0.18	0.13	0.52

Notes: Table 2 presents the results of estimating Equation 5. Independent variable is interacted with dummy for political leaning across all quartiles by policy exposure. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Given the events that precipitated the ban and that white nationalist movement was widely seen as the primary target of the policy change, it is notable that racial and ethnic slurs are the only category for which we do not observe a significant decrease within the

subsample of right-wing users. In general, this suggests that language change that we observe is not driven by self-censorship or response to the political movement of the time as opposed to the policy change itself, since in both cases we would expect larger decreases for the salient categories of hate speech which are racial and ethnic slurs.

In Table C.15 we re-estimate the specifications in Columns (1) and (2) of Table 1 additionally interacting independent variable with an indicator for detailed polarity of a treated user. We see decreases in most subgroups of right-wing users, and the effects are similar in magnitude both in absolute and in relative terms (see Table A.7 for group means). The only exception is users of smaller alt-right forums for which we report similar in relative magnitude but insignificant coefficients.

The null result for alt-right users is likely explained by the fact that, unlike other groups, alt-right forums are, with some exceptions, already dormant at the time of the policy change and we similarly do not observe the effect on volume of activity for them.

Our results in the previous section suggest that substitution that we observe is driven by core left-wing and transphobic users and that at least part of the substitution is driven by migration into relatively more hateful forums. We analyse the effects on hate speech use separately for these groups to see whether the impact of the policy change on them is different in this respect as well.

Table C.9 shows the estimates by quartile of policy exposure and separately for left- and right-wing users. It is clear that the smaller overall effects for core users of banned forums that we have seen previously are a combination of strongly significant effects for *core* right-wing users and positive, if insignificant, effects on *core* left-wing users. Tables C.13 and C.14 show the results of estimating equivalent specifications for *core* left-wing users by category of slur. It is clear that the increase, if any, comes exclusively from higher incidence of disability-related and homophobic language. The increase is relatively small and insignificant when we control for average slur incidence in a forum.

Overall, it seems that migration into relatively more toxic forums might slightly increase slur use of migrating users, but not beyond the average level of the forums-substitutes.

5.4 Spillover effects

In general, our results so far indicate that the policy change leads to a significant reduction in use of slurs by most right-wing treated users. Does that reduction lead to additional spillover effects on non-treated users interacting with the treated?

To get an idea of the magnitude of potential spillovers, we can turn to Columns (1) and (2) in Table C.21 estimating Equation 5 additionally interacted with indicator for whether the comment was a reply to a treated user.[11] Figure C.5 displays the incidence of slurs

[11]The null result in Column (3) likely stems from insufficient power to identify the effect: the share of treated users on the platform is small, so the likelihood of repeated interactions with the same treated user on Reddit is very low.

in replies to treated users by week in the raw data. Baseline incidence of slurs is higher in replies to treated users (see coefficient on $ParentTreated_c$) and is reduced significantly after the policy change (coefficient on $ParentTreated_c \times Post_t$).

The decrease in incidence of hate speech in replies to treated users is not explained by lower overall propensity to reply to slurs with slurs after the policy. In Table C.20 we compare the effect on interactions with indicator for reply to a treated user and reply to a slur. Coefficients on $Post_t \times ParentSlur_c$ are negative but insignificant.

The magnitude of the effect is statistically and economically significant, especially when we take into account that it reflects an absolute change comparable to that of the main policy effect but relative to a lower mean slur incidence.

Thus, the policy change has additional effects over and above influencing use of slurs by users directly affected by the policy through contagion: individual comments are more likely to contain slurs if they are left as a reply to to comment containing slurs. Since treated users use less hate speech after the policy, the replies to their comments become in turn less toxic. As previously, inclusion of user-subreddit fixed effects allows us to rule out that the effect is driven by moderation differences after the policy change.

5.5 Mechanisms

We now turn to the analysis of some of the mechanisms that can drive the reduction in hate speech use of treated users. We consider five main directions: platform-wide changes in *moderation*, *self-censorship* to avoid future bans, *displacement* of hate speech off Reddit, changes in *engagement quality* and reduction of *exposure* of treated users to banned forums.

Changes in moderation We might be worried that the reductions in hate speech are driven not by individual response of treated users, but by wider differences in moderation across the platform.

Moderation on Reddit happens primarily on *subreddit* level subject to local forum rules (see Section 2 for discussion) and is enforced by volunteer moderators who are members of the forum, not Reddit's employees. Hate speech is widely present outside banned forums in the many forums that do not restrict its use. From the raw data displayed in Figure C.1 we do not see significant changes in the overall level of slur incidence on the platform after the policy change that would suggest large scale platform-wide changes in moderation.

Robustness of the main results in specifications with subreddit-, post-, and thread-level fixed effects (Table 1) allows us to rule out that the effect is driven by changes in *subreddit*-level moderation rules: for moderation to be generating our results, it would have to disproportionately affect treated users within the same conversation thread.

There are only two possible scenarios in which such disproportional local moderation can occur. The first is direct targeting of treated users by moderators of individual forums. This appears implausible since treated users are not immediately identifiable as such, unless

they mention it themselves. Given that treated users are only 0.6% of total users of Reddit in our sample and that in most cases moderators do not encounter more than a single treated user, large-scale search and targeting of treated users are not practically feasible.

The second possibility is that there are large differences in composition of slur vocabulary between treated and control group and the specific vocabulary of treated users is affected by post-policy platform-wide moderation differences to a larger extent.

In both cases, the difference in incidence of hate speech would reflect not changes in individual behaviour, but changes in the selection of the comments we observe in the sample due to removals. `Pushshift` data on Reddit allows us to observe moderation process in more detail than is normally feasible for other social media platforms. If comment is deleted by a user or removed by a moderator, we still see it in the data, only without the comment text. Hence, although we do not observe what has been moderated, we still know where the moderation happened.

Self-censorship One plausible explanation is that the decreases in use of slurs that we observe are due to self-censorship if treated users expect to be targeted by subsequent moderation.

Self-censorship in itself is not necessarily a problem if it results in a genuine decrease in hate speech use – that would still mean that the policy is effective, at least in the objectives that are measurable in observational data. The concern is that the easily detectable types of slurs that we measure here are replaced by more veiled forms of hate speech.

Larger emphasis on local moderation and forum-level as opposed to user-level bans means that the incentives for ban evasion on Reddit are different from other major social media platforms. Specifically, we expect attempts at ban evasion to happen not on individual, but on community level. An individual user commenting in a non-toxic forum is not "hiding" from a centralized moderation system flagging potentially hateful content, but from a local moderator, a member of the same forum who has ability to review individual comments within their context. There is little reason to expect that peer *subreddit* members would be less responsive to veiled but locally obvious hate speech than to slurs: hidden toxicity is harder to recognise for an automated system, but not for a human reader. In other words, on user level we do not expect effects on slurs and other types of hate speech to go in different directions.

The only case where there are clear incentives for ban evasion are in forums that might expect to be shut down in subsequent moderation waves. In that case, the reduction in use of slurs would be driven by disproportionate substitution to veiled hate speech by treated users relative to control within "endangered" forums. In that case, we expect larger relative effects for forums perceived to be under threat.

Reddit's management never explained the rationale behind the specific list of forums that were banned in June 2020, and the ban targeted a wide variety of communities in

terms of their topic, size and level of hate speech use. Hence, we do not have any specific criteria that would allow us to identify *subreddits* in danger of a future ban. Instead we consider two plausible proxies: high pre-policy level of slur use and high concentration of treated users.

In Table C.16 we estimate the effects by quartile of pre-policy hate speech use within a subreddit. As quartile of subreddit-level hate speech incidence increases, the absolute magnitude of the coefficient estimates go up. But the effect sizes here have to be interpreted relative to the group means: the baseline slur incidence in forums in fourth quartile is 100 times higher than those in the first one. In relative terms, the magnitude of the coefficients for the first three quartiles is very similar (11 to 13% of the mean). The relative magnitudes for the most toxic forums are very small (0.8%) and insignificant.

There are other empirical facts that are not well explained by self-censorship. Firstly, the effect on use of hate speech is not instantaneous, but gradual and only reaches the maximum level in about two months after the ban (Figure C.7). Forum-level self-censorship is more likely to produce an immediate decrease in slur incidence. Secondly, under self-censorship, we expect larger effects for racial and ethnic slurs, since the policy change was precipitated by the murder of George Floyd and public outcry about the role of Reddit in coordinating white nationalist movement. Instead, racial and ethnic slurs are the only category for which we do *not* observe a significant effect within the sample of right-wing users.

Hence, the idea is to see the effects of the policy on treated and control users in forums similar to banned but not directly affected by the policy change. We see a much larger reduction in use of hate speech by treated users within exposed forums, but when interpreting the results relative to the baseline incidence of slur use within exposed forums, it is actually very similar to the main effect. From descriptive statistics, overall incidence of hate speech in forums with large proportion of treated users is high (although not to the extent of banned ones), and treated users of exposed forums are much more toxic.

Three possible mechanisms are at play here: firstly, users of these forums are definitely aware of the policy change, so a decrease in use of slurs can reflect a degree of self-censorship. Secondly, control users of exposed forums are likely a polluted reference group: many of them were likely active in banned forums we simply to do not observe them in the data, so the effects on them could be just the effects on the treated. Thirdly, since contagion in slur use is likely a thing, the decreases we see could reflect spillover effects.

Displacement of hate speech off Reddit A potential concern is that the ban of hateful forums induces treated users to choose alternative platforms to discuss hateful topics. Then, the reduction in use of slurs that we observe could be the result not of a behavioural change, but of change in composition of content that treated users choose to leave on Reddit. Such compositional change would not affect the evaluation of policy

effectiveness within Reddit, but would put into question desirability of the intervention from policymaker's perspective.

The results that we have presented so far suggest that that displacement off platform is unlikely to be the main driver of the reduction in hate speech. We do not observe mass platform exit or meaningful decrease in overall activity outside banned forums after the policy change. To produce our results, displacement of hate speech off platform would have to compensated with increase in overall number of comments left in the same forums, including *subreddits* with highest incidence of hate speech, since we do not observe a decrease in activity in those forums by right-wing treated users driving the decrease in hate speech use.

Another observation is that after the policy change treated users are still much more toxic than the control group. Additionally, the effects we observe are not uniform across slur types and do not completely correspond to the most prominent slur categories within banned forums either. If displacement were to drive our results, it would have to happen selectively, for slur types that do not align with those of banned forums.

Engagement quality Another possibility is that the ban of hateful forums discourages active platform participation of treated users. Then, the decrease in hate speech could be due to general reduction in comment quality. We construct several proxies of engagement quality to see whether any high-level changes in comment types are detectable in the data. We compute comment length as number of words[12] with the idea that "throwaway" comments (emojis, single-word responses) would be detectable by this simple measure. We construct an indicator for whether comment includes a quote of other user from the same discussion to proxy the degree of involvement in a particular conversation. Additionally, we calculate the number of unique links included in a comment to proxy for informational content. Lastly, we construct the number of direct replies to a comment and number of total replies (the size of the whole conversation tree originating in a comment) as "black box" proxies of the degree of engagement that a comment elicits from other users. The intuition here is that more nuanced changes in the way users talk might be visible in the degree to which other users are willing to engage with them.

[12]Reddit does not impose an upper limit on the number of characters in a comment.

Table 3: Effects on proxies of engagement: difference-in-differences estimates

	Text Length	**# Links**	**# Direct Replies**	**# Total Replies**
	(1)	(2)	(3)	(4)
Post × Treated	1.380***	-0.000	0.004***	0.265
	(0.181)	(0.001)	(0.001)	(0.185)
Fixed effect	**User×Subreddit**	**User×Subreddit**	**User×Subreddit**	**User×Subreddit**
Weeks	53	53	53	53
Groups	241,525,510	241,525,510	241,525,510	241,525,510
Observations	1,852,554,254	1,852,554,254	1,852,554,254	1,852,554,254
Dependent var. mean	137.96	0.069	0.540	5.60

Notes: Table 3 presents the results of estimating Equation 5 with different proxies for engagement quality as dependent variables (the exact specification is equivalent to the one in Column 2 of Table 1). Standard errors in parentheses are clustered at the user level. Column 1 presents the effects on length of comment text (in number of symbols). Column 2 reflects the effect on the number unique links within comment. "Direct Replies" in Column 3 refers to number of comments left as direct replies to a given comment, "Total Replies" in Column 4 additionally includes all replies to the direct replies (that is, the whole conversation tree under a given comment). The results should be taken with caution, as it is clear from Figure C.6 displaying the raw data that pre-trends are unlikely to always hold. Nonetheless, given the small magnitude of the effects relative to variable means, we can conclude that the policy either does not have an effect on the engagement proxies or these effects are incredibly small.

We produce each measure for every comment in the dataset to be able to account for local variation in communication style via forum-, week- and user fixed effects. The estimation strategy is equivalent to the one used to produce results on hate speech use (Equation 5) with forum-by-user fixed effects. Table 3 presents the results of the estimation, Figure C.6 displays raw averages for treatment and control group. We do not see any evidence of negative effects of the policy on engagement quality: if anything, coefficient estimates tend to be positive, although of negligible magnitude.

Exposure Our results so far suggest that plausible indirect effects of the policy, such as differences in selection into different types of forums, overall moderation changes, platform exit or differences in engagement quality cannot fully account for the decreases in hate speech use that we observe. We now consider potential direct effects on language of treated users due to the policy: via norm adjustment, reduced exposure to hateful content or lower level of interactions with the network of other treated users.

Adjustment of individual users to language of the online communities in which they talk is well documented in the literature, for Reddit and other social networking cites (Danilescu et al 2015). Multiple reasons have been suggested, but they can be assigned in one of two main groups. The first is intentional adjustment to the local norms (to signal in-group status, increase communication efficiency, et cetera), the second is "contagion", unintentional influence of the overall language that user is exposed to on the content they produce.

Both of these appear to be likely in this case. Policy change can lead to norm adjustment if being moderated leads user to adjust their perception of what is acceptable on the platform or if wholesale adjustment is needed to successfully communicate in forums substitutes. This effect might be a form of self-censorship, but it is not an empirical concern,

since it still reflects actual decrease in hate speech use, not ban evasion through shift in "veiled" forms of toxicity.

If individual language choice is a function of the whole "diet" of a user, not just the local norm, then the policy change can reduce contagion from banned forums by lowering the overall level of exposure to hate speech. This is the likely channel through which interventions targeting overall exposure to toxicity work in reducing toxicity of users (notably documented by Beknazar-Yuzbashev et al. (2025)), although in our case the contagion would work on hate speech specifically.

We can hardly pin down the individual pathways in observational data, but we can provide suggestive evidence for some of them.

In principle, if the policy change triggers adjustment to local communication norms of non-banned forums or if it works through reducing overall level of exposure to hate speech, we would expect the effects to be less pronounced in forums with high baseline hate speech use. This is what we see in Table C.16 which shows the effects on hate speech use by quartile of forum hate speech level. After adjusting for baseline level of hate speech use by a user in a forum (Columns 2-3), we see a null effect on behaviour of treated users in highly toxic non-banned forums.

Language adjustment is also supported by the observation that we made in Section 5.3 - we observe larger reductions in hate speech use for the categories of slurs in which the difference in average incidence between banned and non-banned forums in which treated user groups are active is the largest.

6 Discussion

To summarise, in this paper we find that the ban of hateful forums on Reddit was successful in reducing hate speech levels on the rest of the platform. The effect comes both from reduction in hate speech use by users directly affected by the ban and through spillover effects on other users. We do not see evidence of substantial backlash or alienation of treated users. At the same time the effects are asymmetric depending on the user group: we see large effects for right-wing users and null for left-wing ones. Results for migration of displaced users suggest that migration is limited to (or is only detectable for) core users of banned forums, predominantly left-wing ones.

The effects we document are economically large, partly due to the fact that the policy impacts suppliers of toxicity with high baseline level of hate speech use and constitutes a substantial shock to individual activity of treated users on Reddit. One could argue that intervention of this type is less relevant from policymakers' standpoint since it directly affects only small number of users. It is worth remembering though that on Reddit, as on other platforms, small number of users account for the vast proportion of overall toxicity levels. Taking into account contagion, moderate differences in language of highly toxic users can have vast multiplicative effects across the platform.

The granularity of the data on user activity on Reddit allows us to shed some light on the mechanisms underlying these results. We document that the effects are unlikely to be driven by self-censorship, platform-level changes in moderation, selection into different conversation types and changes in quality of comments. We document that the use of hate speech on Reddit is predominantly explained by local norms of interaction, not inherent user types and there is significant variation in behaviour of a user and even a user pair across forums that they visit.

We argue that the ban of hateful forums reduces hate speech in the comments of treated users by decreasing the distance between language of affected users and other members of Reddit. We suggest that the reduction can be driven by a variety of underlying behavioural mechanisms: reduced individual exposure to toxicity, conscious adjustment to the norms of the platform or reduction in influence of the network of treated users. Although we cannot support any specific one of them empirically, all of these mechanisms have been documented previously in other settings.

The organization of Reddit as a platform contributes to the effectiveness of the ban since banned communities cannot be easily recreated on the rest of the platform. User networks on Reddit are brittle: individual treated users who interact often can coordinate after the ban, but the bulk of a banned forum cannot be recreated easily, since there is no obvious coordination mechanism.

Overall, the same features of Reddit that make enforcing deplatforming of individual users hard – anonymity, ease of account creation and lack of reliance on offline user networks – are also the features that make the ban of toxic forums effective, since they reduce the

degree to which a user is "entrenched" in a specific mode of social media consumption.

In this sense, Reddit is dissimilar to most other major social media platforms where users are largely connected by personal profiles, not their interest groups. A ban of hateful forums is a simultaneous shock to the network of toxic users and to the degree of individual exposure to hate speech on a scale that would be hard to reproduce on Facebook, Twitter/X or TikTok. At the same time, there is no reason to suspect that the underlying mechanisms driving policy response are different across the types of moderation interventions. Most notably, bans of high-profile toxic users likewise reduce exposure of their active followers to hateful content, remove a digital space where hate speech is normalized and weaken the network of users interacting in the comment sections of those users.[13]

We additionally highlight some empirical patterns which so far have been largely overlooked by economic literature. First, we draw attention to heterogeneity in individual response by characteristics of targeted users, likely driven by costs of substitution within and outside a platform. Differential costs of substitution can underlie not only differences between groups of users within platforms, but also differences in effects between general interest and alternative platforms documented by previous work. Second, we document high degree of variation in individual behaviour depending on local conversational norm: we suggest that the effectiveness of any moderation intervention has to be judged by analysing individual response relative to the "norm space" in which user is active and quantifying the extent to which an intervention works through inducing a language change versus changing the type of conversations users chose to engage in. We hope that these insights can help inform causal analysis of moderation interventions in other settings.

[13]Even on user-based social media like Facebook and X, part of the disproportional influence of high profile toxic users is not only their reach and persuasiveness, but the fact that their *comment sections* provide space for large amounts of highly concentrated and "locally normalized" toxic content. Following a single toxic user can massively increase effective exposure to toxicity when we take into account the amount of *replies* that their posts get.

7 Conclusion

This paper examines the effects of restricting toxic online spaces on the behaviour of affected users. Overall, we find that the most active users of banned *subreddits* post more, move to new forums and use hate speech less after the ban of forums promoting hateful content. Our results indicate the policy had beneficial spillover effects on the toxicity of users interacting with those affected by the policy. We do not find evidence of large scale platform exit, re-creation of banned communities elsewhere on Reddit or alienation of affected users.

We conclude that, from platform's perspective, the moderation policy was effective, since the shutdown of toxic forums induced reduction in hate speech use outside banned forums without repercussions on degree of engagement of moderated users. We document that the behaviour of users on Reddit is largely governed by community-level conversational norms, as opposed to individual characteristics, and the reductions in hate speech are driven by the language types in which distance between banned forums and the rest of Reddit was the largest. Overall, our results point to the conclusion that the policy change works since it can induce large user-level reductions in exposure to norms and language of hateful forums. At the same time, we document substantial heterogeneity in effect across users, highlighting that moderation policies might have unintentionally uneven, context dependent effects across targeted groups.

8 Ethical considerations

The analysis in this paper is based entirely on publicly available user-generated content. No attempt has been made to collect data deleted by the users. Care has been taken to ensure that no results presented in this paper can be used to de-anonymise or target individual users: all figures and exhibits use aggregated data based on no fewer than 500 accounts. We do not include the names of banned forums with fewer than 10 daily users. Furthermore, we anonymise users in the data used for empirical analysis using irreversible hash encoding. No user-generated data has been passed on to third party applications, including AI tools. The study was approved by the Humanities and Social Sciences Research Ethics Committee at the University of Warwick.

References

Agarwal, Saharsh, Uttara M Ananthakrishnan, and Catherine E Tucker (2022). "Content moderation at the infrastructure layer: Evidence from parler". *SSRN Electronic Journal*.

Aliapoulios, Max, Emmi Bevensee, Jeremy Blackburn, Barry Bradlyn, Emiliano De Cristofaro, Gianluca Stringhini, and Savvas Zannettou (2021). "A large open dataset from the Parler social network". *Proceedings of the International AAAI Conference on Web and Social Media*. Vol. 15, pp. 943–951.

Allcott, Hunt, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow (2020). "The welfare effects of social media". *American Economic Review* 110.3, pp. 629–676.

Allcott, Hunt and Matthew Gentzkow (2017). "Social media and fake news in the 2016 election". *Journal of economic perspectives* 31.2, pp. 211–236.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song (2024). "The economics of social media". *Journal of Economic Literature* 62.4, pp. 1422–1474.

Baumgartner, Jason, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn (2020). "The Pushshift Reddit Dataset". *Proceedings of the international AAAI conference on web and social media*. Vol. 14, pp. 830–839.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski (2025). *Toxic content and user engagement on social media: Evidence from a field experiment*. Tech. rep. CESifo Working Paper.

Braghieri, Luca, Ro'ee Levy, and Alexey Makarin (2022). "Social media and mental health". *American Economic Review* 112.11, pp. 3660–3693.

Callaway, Brantly, Andrew Goodman-Bacon, and Pedro HC Sant'Anna (2024). *Difference-in-differences with a continuous treatment*. Tech. rep. National Bureau of Economic Research.

Cao, Andy, Jason M Lindo, and Jiee Zhong (2023). "Can social media rhetoric incite hate incidents? Evidence from Trump's "Chinese Virus" tweets". *Journal of Urban Economics* 137, p. 103590.

Chandrasekharan, Eshwar, Shagun Jhaver, Amy Bruckman, and Eric Gilbert (2022). "Quarantined! Examining the effects of a community-wide moderation intervention on Reddit". *ACM Transactions on Computer-Human Interaction (TOCHI)* 29.4, pp. 1–26.

Chandrasekharan, Eshwar, Umashanthi Pavalanathan, Anirudh Srinivasan, Adam Glynn, Jacob Eisenstein, and Eric Gilbert (2017). "You can't stay here: The efficacy of reddit's 2015 ban examined through hate speech". *Proceedings of the ACM on human-computer interaction* 1.CSCW, pp. 1–22.

Chen, Jiafeng and Jonathan Roth (2024). "Logs with zeros? Some problems and solutions". *The Quarterly Journal of Economics* 139.2, pp. 891–936.

Cima, Lorenzo, Benedetta Tessa, Amaury Trujillo, Stefano Cresci, and Marco Avvenuti (2025). "Investigating the heterogeneous effects of a massive content moderation intervention via Difference-in-Differences". *Online Social Networks and Media* 48, p. 100320.

Cima, Lorenzo, Amaury Trujillo, Marco Avvenuti, and Stefano Cresci (2024). "The Great Ban: Efficacy and Unintended Consequences of a Massive Deplatforming Operation on Reddit". *Companion Publication of the 16th ACM Web Science Conference*, pp. 85–93.

Duong, Thuy (Nov. 19, 2025a). *Most popular social networks worldwide as of February 2025, by number of monthly active users (Statista)*. https://www.statista.com/statistics/272014/global-social-networks-ranked-by-number-of-users/. Accessed: 26-02-2026.

— (Nov. 19, 2025b). *Most popular websites worldwide as of August 2025, by total visits (Statista)*. https://www.statista.com/statistics/1201880/most-visited-websites-worldwide/. Accessed: 26-02-2026.

— (Aug. 4, 2026a). *Average number of daily active users (DAU) on Reddit from 1st quarter 2021 to 2nd quarter 2026, by online status (Statista)*. https://www.statista.com/statistics/1453133/reddit-quarterly-dau-by-online-status/. Accessed: 10-09-2026.

— (July 1, 2026b). *Average number of daily active users (DAUs) on Reddit from 1st quarter 2021 to 2nd quarter 2026, by region (Statista)*. https://www.statista.com/statistics/1453149/reddit-quarterly-dau-by-region/. Accessed: 10-09-2026.

Ershov, Daniel and Juan S Morales (2024). "Sharing News Left and Right: Frictions and Misinformation on Twitter". *The Economic Journal*, ueae027.

Gauthier, Germain, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya (2026). "The political effects of X's feed algorithm". *Nature*, pp. 1–8.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. (2023). "How do social media feed algorithms affect attitudes and behavior in an election campaign?" *Science* 381.6656, pp. 398–404.

Honnibal, Matthew, Ines Montani, Sofie Van Landeghem, Adriane Boyd, et al. (2020). "spaCy: Industrial-strength natural language processing in Python".

Horta Ribeiro, Manoel, Justin Cheng, and Robert West (2023a). "Automated content moderation increases adherence to community guidelines". *Proceedings of the ACM web conference 2023*, pp. 2666–2676.

Horta Ribeiro, Manoel, Homa Hosseinmardi, Robert West, and Duncan J Watts (2023b). "Deplatforming did not decrease Parler users' activity on fringe social media". *PNAS nexus* 2.3, pgad035.

Horta Ribeiro, Manoel, Shagun Jhaver, Savvas Zannettou, Jeremy Blackburn, Gianluca Stringhini, Emiliano De Cristofaro, and Robert West (2021). "Do platform migrations compromise content moderation? evidence from r/the_donald and r/incels". *Proceedings of the ACM on Human-Computer Interaction* 5.CSCW2, pp. 1–24.

Huffman, Steve (Aug. 2015). *Content Policy Update*. Accessed: 10-09-2026. Reddit. URL: https://www.reddit.com/r/announcements/comments/3fx2au/comment/ctsqobs/ (visited on 08/05/2015).

— (Oct. 2019). *Testimony Submitted for the Record: Fostering a Healthier Internet to Protect Consumers*. Written testimony, U.S. House of Representatives, Committee on

Energy and Commerce, Subcommittees on Communications & Technology and Consumer Protection & Commerce. 116th Congress.

Huffman, Steve (June 2020a). *Upcoming changes to our content policy, our board, and where we're going from here*. Accessed: 10-09-2026. Reddit. URL: https://www.reddit.com/r/announcements/comments/gxas21/upcoming_changes_to_our_content_policy_our_board/ (visited on 06/05/2020).

— (June 2020b). *Update to Our Content Policy*. Accessed: 10-09-2026. Reddit. URL: https://www.reddit.com/r/announcements/comments/hi3oht/update_to_our_content_policy/ (visited on 06/29/2020).

Jiménez-Durán, Rafael (2023). "The economics of content moderation: Theory and experimental evidence from hate speech on Twitter". *George J. Stigler Center for the Study of the Economy & the State Working Paper* 324.

Jiménez-Durán, Rafael, Karsten Müller, and Carlo Schwarz (2023). "The Effect of Content Moderation on Online and Offline Hate: Evidence from Germany's NetzDG". *Available at SSRN 4230296*.

Kennedy, Chris J, Geoff Bacon, Alexander Sahn, and Claudia von Vacano (2020). "Constructing interval variables via faceted Rasch measurement and multitask deep learning: a hate speech application". *arXiv preprint arXiv:2009.10277*.

Lee, David S (2009). "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects". *The Review of Economic Studies* 76.3, pp. 1071–1102.

Levy, Ro'ee (2021). "Social media, news consumption, and polarization: Evidence from a field experiment". *American economic review* 111.3, pp. 831–870.

McRae, Michael (2025). "Vertical Governance of Online Speech: Evidence from Google's Moderation Mandate". *Available at RePEc*.

Müller, Karsten and Carlo Schwarz (Oct. 2020). "Fanning the Flames of Hate: Social Media and Hate Crime". *Journal of the European Economic Association* 19.4, pp. 2131–2167. ISSN: 1542-4766.

— (2023). "From Hashtag to Hate Crime: Twitter and Antiminority Sentiment". *American Economic Journal: Applied Economics* 15.3, pp. 270–312.

Ofcom (2026). *Experiences of hateful content*. Report. Accessed: 26-09-2026. Office of Communications (Ofcom).

r/AgainstHateSubreddits (June 2020). *Open Letter to Steve Huffman and the Board of Directors of Reddit, Inc – If you believe in standing up to hate and supporting black lives, you need to act*. Accessed: 10-09-2026. Reddit. URL: https://www.reddit.com/r/AgainstHateSubreddits/comments/gyyqem/open_letter_to_steve_huffman_and_the_board_of/ (visited on 06/08/2020).

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

Statista (2024). *Ranking of the number of Reddit users by country 2024*. https://www.statista.com/forecasts/1174696/reddit-user-by-country/. Accessed: 10-09-2026.

Thomas, Daniel Robert and Laila A Wahedi (2023). "Disrupting hate: The effect of deplatforming hate organizations on their online audience". *Proceedings of the National Academy of Sciences* 120.24, e2214080120.

Trujillo, Amaury and Stefano Cresci (2022). "Make reddit great again: assessing community effects of moderation interventions on r/the_donald". *Proceedings of the ACM on Human-computer Interaction* 6.CSCW2, pp. 1–28.

Appendix

A Descriptive statistics . . . 52
- Activity on the platform . . . 52
- Treated users . . . 52
- Use of hate speech . . . 56

B Additional results on displacement and migration . . . 65
- Raw data . . . 65
- Migration . . . 69
- Alternative specifications & user heterogeneity . . . 78

C Additional results on language change . . . 94
- Raw data . . . 94
- Robustness to alternative estimations . . . 97
- Heterogeneity by user group . . . 104
- Heterogeneity by local language norm . . . 113
- Comment toxicity . . . 120

D Platform overview . . . 123

A Descriptive statistics

Activity on the platform

Table A.1: Descriptive statistics

	Mean	SD	Min	Median	Max
Subreddits (n=937,450)					
# weeks present	6.360	12.074	1	1	53
Average # of comments per week	49.268	2,300.894	1	2	1,559,300.189
Average # of users per week	15.043	561.810	1	1	380,587.906
Banned subreddits (n=112)					
# weeks present	11.080	4.454	1	14	14
Average # of comments per week	2,793.997	15,543.876	1	58.5	124,657.071
Average # of users per week	389.899	1,571.955	1	24.893	11,244.5
Users (n=29,406,985)					
# of comments per week	63.684	486.155	1	5	894,963
Average # forums visited in a week	2.682	4.154	1	1	4,968
Treated users (n=189,950)					
# of weeks present	9.216	4.526	1	10	14
# of comments	19.052	47.654	1	6	17,846
Avg # forums visited in a week	6.505	13.562	1	3	4,968
# of banned subreddits used	1.202	0.743	1	1	36

Notes: The table presents descriptive statistics for the near population of comments on Reddit in 2020. The statistics include comments left by bots or deleted accounts. User-level statistics exclude both categories. In total, we observe 2,161,449,965 comments left on the platform in 2020. 10.6% of comments in the database are missing author id: these are removed comments or comments left by deleted accounts.

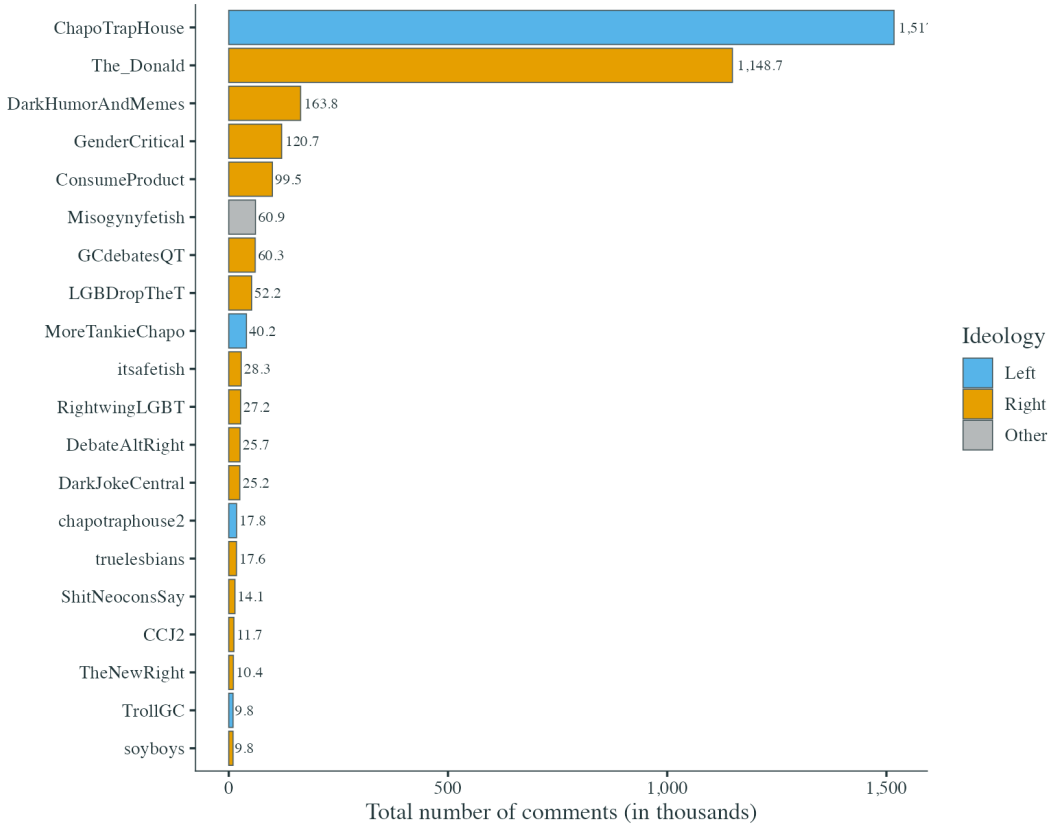
Treated users

Table A.2: Descriptive statistics: volume of activity of treated users by political leaning

	N	Mean	SD	Min	Median	Max
Panel A: Subreddits						
Average weekly # of comments						
Left	7	18649.412	46761.276	1	856.428	124657.072
Right	95	1850.096	11195.120	1	50.384	107788.0
Other	10	662.271	1768.312	7.428	91.214	5689.285
Average weekly # of users						
Left	7	1852.404	4100.603	1	273.857	11117.857
Right	95	303.074	1287.456	1	23.785	11244.50
Other	10	190.978	408.602	5.642	57.928	1345.642
Panel B: Users						
# of comments						
Left	50,179	17.211	42.353	1	6	4174
Right	137,050	19.279	44.009	1	7	17846
Other	2,721	38.448	154.029	1	13	11230

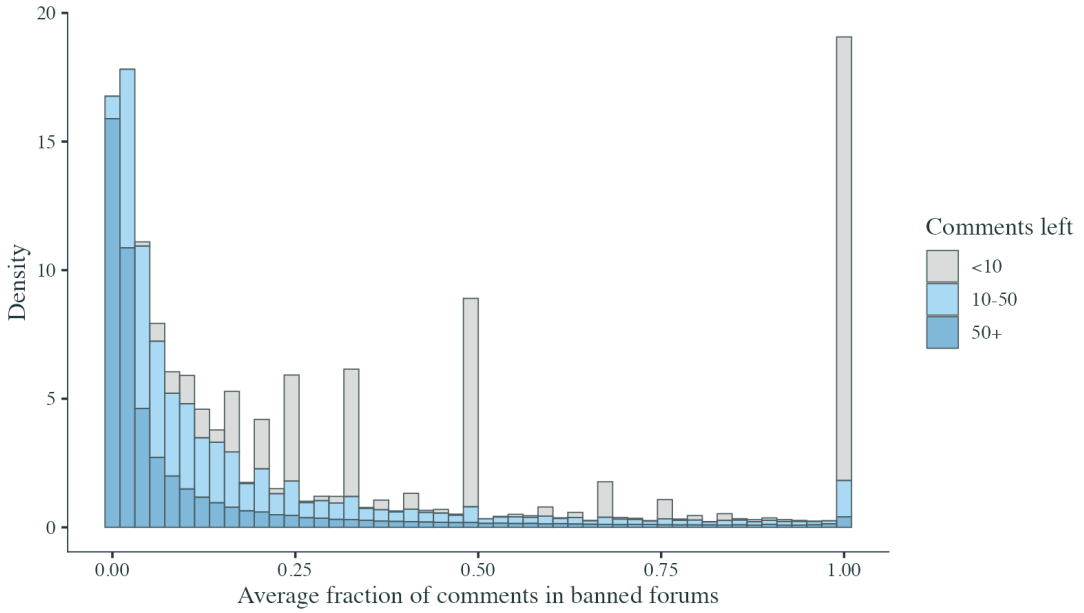
Notes: The table presents descriptive statistics showing volume of activity in banned subreddits by subreddit type (Panel A), as well as volume of activity of treated users depending on whether they post in predominantly left- or right-leaning banned subreddits (Panel B). User is classified as right- or left- leaning if 95% of their activity occurs in the corresponding subreddit type (or in banned subreddits classified as "Other"). Less than 1% of treated users comment in both left- and right-leaning forums.

Figure A.1: Top-20 largest banned subreddits by ideology



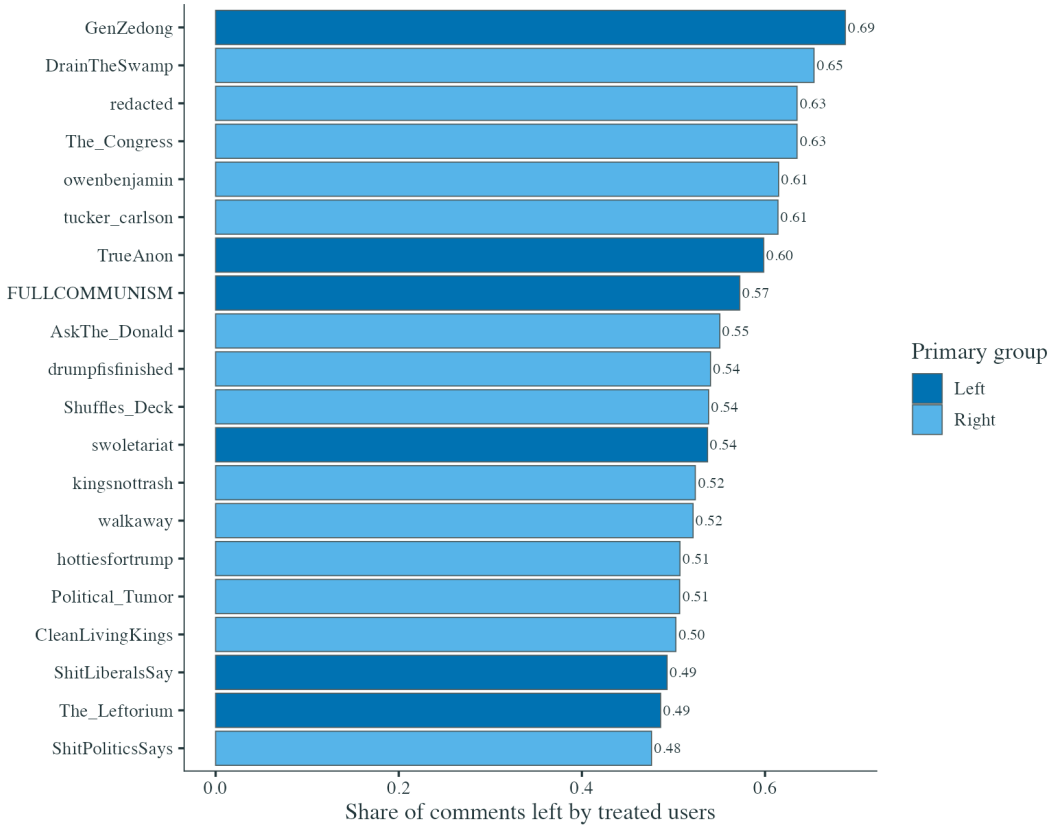
Notes: Figure A.1 shows total number of comments (in thousands) left in 20 largest banned subreddits in January-March 2020. Banned left-wing forums denoted in blue, banned right-wing forums denoted in orange. From 112 identified banned subreddits, 95 can be classified as right- wing, 7 as left-wing. Additionally, 10 forums belong in neither category, primarily containing extreme misogynistic sexual content.

Figure A.2: Distribution of policy exposure across treated users



Notes: Figure A.2 shows the distribution of policy exposure across treated users. Exposure ($ShareInBan_i$) is measured as the fraction of total comments left in subsequently banned subreddits by week, averaged over January-March 2020. Color denotes three subgroups of users based on the total number of comments left on Reddit in January-March 2020. Average over weeks is preferred to simple average over the period to put larger relative weight on repeated engagement in banned forums. For reasons discussed in Section 4.1, average exposure does not take into account zeroes for weeks in which user did not leave any comments on the platform. The Figure illustrates that exposure to the policy measured as a share of total activity leads to oversampling of users who comment little overall in higher quartiles of exposure. For this reason, we ensure that all main results in the paper are robust to estimation on subsample of users who leave at least 10 or 50 comments between January and March 2020.

Figure A.3: Top 20 non-banned subreddits with largest share of treated users



Notes: Figure A.3 plots top 20 non-banned subreddits with the highest share of comments left by users of banned subreddits. The sample is restricted to subreddits with over 5000 comments left between January and June 2020. The color denotes the primary group of treated users active in the corresponding subreddit by ideology.

Use of hate speech

Table A.3: Incidence of slurs: banned vs non-banned forums

	Non-banned				Banned			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Number of slurs per comment								
All categories	2.526	41.728	0	21,000	17.411	23.286	0	119
Antisemitic	0.014	3.093	0	1,500	0.449	1.763	0	14
Disability	0.285	10.501	0	1,000	3.169	8.476	0	62
Homophobic	0.323	14.251	0	7,000	3.623	7.474	0	50
Misogynistic	1.119	23.907	0	3,000	4.955	13.231	0	102
Racial/ethnic	0.732	21.173	0	9,000	3.939	12.483	0	99
Transphobic	0.052	4.726	0	2,000	1.277	7.886	0	83
Share of comments containing slurs								
All categories	2.368	32.461	0	1,000	16.246	20.988	0	114
Antisemitic	0.013	2.727	0	1,000	0.445	1.760	0	14
Disability	0.284	10.463	0	1,000	3.159	8.474	0	62
Homophobic	0.310	11.938	0	1,000	3.516	7.307	0	50
Misogynistic	1.064	21.754	0	1,000	4.745	12.575	0	95
Racial/ethnic	0.691	17.708	0	1,000	3.635	11.126	0	99
Transphobic	0.051	4.351	0	1,000	1.276	7.886	0	83

Notes: Table A.3 shows the descriptive statistics on the incidence of hate speech by language type separately for banned and non-banned forums. Incidence of slur use is measured as the average number of unique words of a given category per comment left in a subreddit (upper panel) or the average number of comments containing slurs of a given category (lower panel). The unit of observation is this a *forum*. High maximum values in the non-banned group are driven by small forums with few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text. The Figure clearly shows that (a) average incidence of slurs is much higher in banned forums, and (b) the difference between estimates using number of unique slurs and dummy whether a comment contains a slur are very similar, that is, the variation is primarily driven by extensive margin of slur use. $N_{banned} = 112, N_{non-banned} = 937,339$.

Table A.4: Incidence of slurs by type and orientation of the banned forum

	All types	*Antisemitic*	*Disability*	*Homophobic*	*Misogynistic*	*Racial/ethnic*	*Transphobic*
Number of slurs per comment							
Right	17.011	0.533	3.610	3.988	2.815	4.555	1.509
Left	6.856	0.009	1.697	3.546	0.480	1.097	0.028
Other	30.343	0.001	0.327	0.567	28.912	0.463	0.072
Share of comments containing slurs							
Right	15.866	0.529	3.604	3.862	2.789	4.198	1.508
Left	6.715	0.009	1.688	3.545	0.442	1.035	0.028
Other	28.152	0.001	0.278	0.567	26.810	0.462	0.072

Notes: Table A.4 shows the descriptive statistics on the incidence of slurs by language type depending on the political orientation of a banned forum. Incidence of slur use is measured as the average number of unique words of a given category per comment within a forum. The unit of observation is thus a *forum*. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text. The Figure illustrates the primary slur types that treated users of corresponding banned forums are exposed to.

Table A.5: Incidence of slurs: treated vs control users

	N	Mean	SD	Min	Max
All categories					
Control	29,216,533	1.601	27.969	0	42,000
Treated	189,946	5.823	31.416	0	2,000
Left	36,364	2.290	13.545	0	1,000
Right	137,213	4.779	24.551	0	2,000
Antisemitic					
Control	29,216,533	0.009	1.981	0	3,000
Treated	189,946	0.039	1.359	0	250
Left	36,364	0.025	1.150	0	200
Right	137,213	0.045	1.471	0	250
Disability					
Control	29,216,533	0.228	8.682	0	2,000
Treated	189,946	1.065	8.852	0	1,000
Left	36,364	0.629	6.029	0	500
Right	137,213	1.257	9.845	0	1,000
Homophobic					
Control	29,216,533	0.173	8.942	0	14,000
Treated	189,946	0.862	10.231	0	1,000
Left	36,364	0.384	6.799	0	1,000
Right	137,213	1.030	11.313	0	1,000
Misogynistic					
Control	29,216,533	0.706	17.435	0	6,000
Treated	189,946	2.680	23.722	0	1,500
Left	36,364	0.595	6.657	0	1,000
Right	137,213	1.155	9.789	0	1,000
Racial/ethnic					
Control	29,216,533	0.447	15.159	0	18,000
Treated	189,946	0.998	12.231	0	2,000
Left	36,364	0.577	5.606	0	500
Right	137,213	1.078	13.607	0	2,000
Transphobic					
Control	29,216,533	0.038	4.008	0	3,000
Treated	189,946	0.179	4.734	0	1,000
Left	36,364	0.080	2.592	0	407
Right	137,213	0.213	5.345	0	1,000

Notes: Table A.5 shows the descriptive statistics on the use of slurs by user group and language type. Incidence of slur use is measured as the average number of unique words of a given category per comment of a user. The unit of observation is thus a *user*. High maximum values are driven by users who left few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text. The number of observations for left- and right-wing users separately does not sum up to the total number of treated users due to the exclusion of treated users not primarily active in political right- or left-wing forums. The Figure shows that the incidence of slurs is much higher for treated users across all categories. With the exception of misogynistic language by left-wing users, the same holds true for right- and left-wing users separately as well, although the values for the right-wing users are much larger.

Table A.6: Use of slurs by treated users inside vs outside banned forums

	Non-banned				Banned			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Number of slurs per comment								
All categories	4.827	25.564	0	2,000	12.865	91.026	0	5,000
Antisemitic	0.036	1.434	0	250	0.073	5.613	0	1,000
Disability	0.974	7.239	0	1,000	1.740	29.976	0	2,000
Homophobic	0.652	7.655	0	1,000	1.991	31.468	0	2,000
Misogynistic	2.149	19.010	0	2,000	6.641	68.219	0	3,000
Racial/ethnic	0.897	10.297	0	1,667	1.955	35.156	0	3,000
Transphobic	0.119	2.695	0	423	0.466	14.192	0	1,000
Share of comments containing slurs								
All categories	4.550	22.307	0	1,000	12.095	82.376	0	1,000
Antisemitic	0.035	1.359	0	250	0.072	5.591	0	1,000
Disability	0.970	7.228	0	1,000	1.716	28.891	0	1,000
Homophobic	0.638	7.545	0	1,000	1.932	30.172	0	1,000
Misogynistic	2.036	16.603	0	1,000	6.255	62.564	0	1,000
Racial/ethnic	0.847	8.870	0	1,000	1.844	31.687	0	1,000
Transphobic	0.116	2.532	0	333	0.465	14.191	0	1,000

Notes: Table A.6 shows the descriptive statistics on the use of slurs by treated users and language type separately for banned and non-banned forums. Incidence of slur use is measured as the average number of unique words of a given category per comment of a user (upper panel) or the average number of comments containing slurs of a given category (lower panel). The unit of observation is thus a *user*. High maximum values are driven by users who left few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text. The Figure shows that treated users use a lot more slurs within banned forums: the mean slur incidence within banned forums is in most cases *at least* double that of non-banned. The observation reinforces the argument that use of hate speech on Reddit is driven by norms of interaction *within a community*, not by invidivual users uniformly more toxic across the platform.

Table A.7: Incidence of slurs by slur category and treated user group

	All types	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
Left							
Non-banned	2.415	0.024	0.615	0.356	0.680	0.664	0.075
Banned	2.778	0.038	0.790	0.669	0.557	0.628	0.095
Right							
Non-banned	4.291	0.042	1.142	0.797	1.160	1.006	0.145
Banned	9.126	0.086	2.099	2.482	1.427	2.429	0.603
Alt-right							
Non-banned	7.415	0.136	2.097	1.476	1.411	2.060	0.234
Banned	15.673	0.255	4.707	3.871	2.200	3.943	0.697
Offensive							
Non-banned	3.121	0.011	0.703	0.496	1.009	0.857	0.046
Banned	9.140	0.087	1.758	1.766	1.347	3.896	0.286
TheDonald							
Non-banned	3.659	0.027	1.171	0.710	1.039	0.567	0.146
Banned	6.522	0.003	1.578	2.596	0.979	0.430	0.936
Transphobic							
Non-banned	3.299	0.017	0.440	0.554	1.491	0.502	0.295
Banned	3.420	0.006	0.054	0.992	1.793	0.284	0.292

Notes: Table A.7 shows the descriptive statistics on the use of slurs by user group and slur type inside and outside banned forums. Incidence of slur use is measured as the average number of unique words of a given category per comment of a user. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text.

Table A.8: Use of slurs by treated users inside vs outside banned forums, by quartile of policy exposure

	Non-banned				Banned			
	Mean	SD	Min	Max	Mean	SD	Min	Max
All categories								
Q1	3.852	10.681	0	811	12.045	97.548	0	3,000
Q2	4.723	16.854	0	919	13.423	91.680	0	5,000
Q3	5.553	26.338	0	1,500	13.894	87.957	0	3,000
Q4	5.312	41.719	0	2,000	12.128	86.301	0	2,000
Antisemitic								
Q1	0.022	0.380	0	41	0.075	7.139	0	1,000
Q2	0.029	0.897	0	153	0.034	2.153	0	250
Q3	0.054	1.885	0	250	0.118	7.923	0	1,000
Q4	0.043	2.058	0	250	0.066	3.029	0	333
Disability								
Q1	0.993	3.508	0	172	1.813	36.807	0	2,000
Q2	1.057	4.861	0	300	2.118	32.240	0	1,000
Q3	1.011	6.325	0	333	1.677	25.237	0	2,000
Q4	0.803	12.381	0	1,000	1.340	23.354	0	1,000
Homophobic								
Q1	0.505	2.562	0	200	1.919	35.694	0	1,000
Q2	0.628	4.321	0	262	2.072	33.322	0	2,000
Q3	0.779	7.841	0	889	1.956	26.681	0	1,000
Q4	0.714	13.046	0	1,000	2.013	29.140	0	2,000
Misogynistic								
Q1	1.508	7.010	0	797	5.884	69.723	0	2,000
Q2	2.011	12.368	0	919	6.422	63.185	0	2,000
Q3	2.478	18.777	0	1,500	7.835	72.005	0	3,000
Q4	2.732	31.968	0	2,000	6.478	67.875	0	2,000
Racial/ethnic								
Q1	0.737	4.663	0	640	2.033	38.624	0	2,000
Q2	0.894	6.178	0	500	2.294	38.744	0	3,000
Q3	1.068	10.602	0	846	1.767	28.645	0	2,000
Q4	0.896	16.993	0	1,667	1.710	33.268	0	2,000
Transphobic								
Q1	0.087	0.853	0	53	0.322	15.067	0	1,000
Q2	0.104	1.286	0	100	0.482	14.540	0	1,000
Q3	0.163	3.868	0	423	0.542	12.979	0	1,000
Q4	0.124	3.649	0	333	0.521	14.040	0	1,000

Notes: Table A.8 shows the descriptive statistics on the use of slurs by treated users and language type separately for banned and non-banned forums and by quartile of policy exposure ($ShareInBan_i$). Incidence of slur use is measured as the average number of unique words of a given category per comment of a user. The unit of observation is thus a *user*. High maximum values are driven by users who left few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text.

Table A.9: Use of slurs by treated left-wing users inside vs outside banned forums, by quartile of policy exposure

	Non-banned				Banned			
	Mean	SD	Min	Max	Mean	SD	Min	Max
All categories								
Q1	2.854	7.813	0	248	4.259	64.951	0	2,000
Q2	2.445	7.934	0	192	4.169	54.046	0	2,000
Q3	2.157	9.867	0	462	2.300	29.986	0	1,000
Q4	2.258	27.321	0	2,000	1.822	26.641	0	1,000
Antisemitic								
Q1	0.023	0.285	0	9	0.000	0.000	0	0
Q2	0.028	0.840	0	67	0.007	0.642	0	56
Q3	0.045	2.437	0	250	0.104	6.276	0	500
Q4	0.007	0.241	0	20	0.016	0.786	0	83
Disability								
Q1	0.864	3.590	0	110	1.436	31.517	0	1,000
Q2	0.720	3.724	0	90	1.388	28.190	0	1,000
Q3	0.613	4.312	0	143	0.627	15.752	0	1,000
Q4	0.472	8.060	0	500	0.341	8.391	0	500
Homophobic								
Q1	0.500	3.771	0	200	1.699	39.203	0	1,000
Q2	0.374	2.910	0	111	0.575	16.553	0	1,000
Q3	0.330	3.287	0	167	0.676	18.088	0	1,000
Q4	0.265	4.665	0	333	0.329	12.726	0	1,000
Misogynistic								
Q1	0.717	2.064	0	34	0.485	20.445	0	1,000
Q2	0.650	3.458	0	167	0.809	20.333	0	1,000
Q3	0.521	3.267	0	143	0.432	11.456	0	1,000
Q4	0.766	20.269	0	2,000	0.543	14.032	0	1,000
Racial/ethnic								
Q1	0.657	3.081	0	111	0.435	16.275	0	1,000
Q2	0.598	3.190	0	107	1.237	31.296	0	2,000
Q3	0.512	3.703	0	148	0.378	8.609	0	500
Q4	0.706	14.658	0	1,000	0.561	16.488	0	1,000
Transphobic								
Q1	0.094	1.098	0	53	0.204	14.270	0	1,000
Q2	0.075	1.248	0	62	0.152	11.615	0	1,000
Q3	0.136	4.858	0	423	0.084	5.178	0	500
Q4	0.044	1.006	0	62	0.031	2.231	0	250

Notes: Table A.9 shows the descriptive statistics on the use of slurs by treated left-wing users and language type separately for banned and non-banned forums and by quartile of policy exposure ($ShareInBan_i$). Incidence of slur use is measured as the average number of unique words of a given category per comment of a user. The unit of observation is thus a *user*. High maximum values are driven by users who left few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text.

Table A.10: Use of slurs by treated right-wing users inside vs outside banned forums, by quartile of policy exposure

	Non-banned				Banned			
	Mean	SD	Min	Max	Mean	SD	Min	Max
All categories								
Q1	3.492	8.470	0	649	8.260	77.858	0	2,000
Q2	4.088	12.651	0	500	9.702	77.302	0	5,000
Q3	4.676	20.643	0	1,026	9.374	62.851	0	2,000
Q4	4.386	37.002	0	2,000	9.259	68.427	0	2,000
Antisemitic								
Q1	0.023	0.405	0	41	0.083	7.810	0	1,000
Q2	0.031	0.956	0	153	0.044	2.464	0	250
Q3	0.057	1.715	0	162	0.137	8.865	0	1,000
Q4	0.064	2.602	0	250	0.087	3.619	0	333
Disability								
Q1	1.084	3.611	0	172	1.997	38.215	0	2,000
Q2	1.190	5.210	0	300	2.397	34.357	0	1,000
Q3	1.224	7.137	0	333	2.113	26.678	0	1,000
Q4	1.038	14.601	0	1,000	1.860	28.032	0	1,000
Homophobic								
Q1	0.514	2.392	0	138	2.051	36.406	0	1,000
Q2	0.697	4.588	0	262	2.559	37.429	0	2,000
Q3	0.957	9.077	0	889	2.564	29.908	0	1,000
Q4	0.951	14.853	0	1,000	2.835	34.252	0	2,000
Misogynistic								
Q1	1.043	2.795	0	134	1.361	32.270	0	2,000
Q2	1.147	4.898	0	150	1.434	26.015	0	1,000
Q3	1.139	7.674	0	382	1.450	25.362	0	2,000
Q4	1.217	22.224	0	2,000	1.473	26.559	0	1,000
Racial/ethnic								
Q1	0.743	4.728	0	640	2.389	42.106	0	2,000
Q2	0.911	6.211	0	500	2.666	41.656	0	3,000
Q3	1.116	11.041	0	846	2.352	33.875	0	2,000
Q4	0.959	18.265	0	1,667	2.280	39.237	0	2,000
Transphobic								
Q1	0.086	0.805	0	53	0.379	16.107	0	1,000
Q2	0.112	1.317	0	100	0.601	15.882	0	1,000
Q3	0.183	3.684	0	333	0.758	15.418	0	1,000
Q4	0.157	3.975	0	308	0.724	16.026	0	1,000

Notes: Table A.10 shows the descriptive statistics on the use of slurs by treated right-wing users and language type separately for banned and non-banned forums and by quartile of policy exposure ($ShareInBan_i$). Incidence of slur use is measured as the average number of unique words of a given category per comment of a user. The unit of observation is thus a *user*. High maximum values are driven by users who left few highly toxic comments. Comments left by bots and users whose account has been deleted are excluded from the sample. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text.

Table A.11: Incidence of hate speech by level of user exposure to slur type

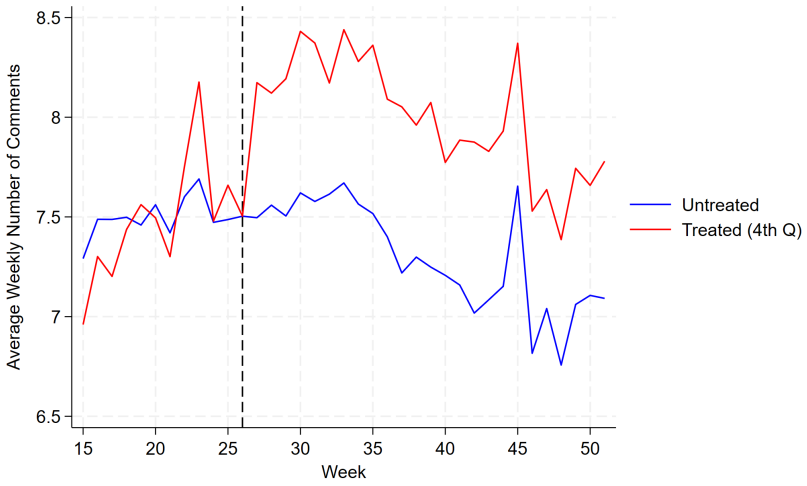
		Antisemitic	*Disability*	*Homophobic*	*Misogynistic*	*Racial/ethnic*	*Transphobic*
Non-banned	Low	0.029	0.509	0.389	0.863	0.614	0.085
	High	0.135	2.080	1.530	6.968	0.864	0.205
Banned	Low	0.003	0.489	0.566	0.652	0.399	0.060
	High	0.330	4.984	8.595	30.141	4.970	0.789
Treated		0.039	1.065	0.862	2.680	0.998	0.179
Control		0.009	0.228	0.173	0.706	0.447	0.038

Notes: Table A.11 shows the descriptive statistics on the incidence hate speech in comments of treated users by slur type depending degree of user exposure (high vs low) to specific type of hate speech within banned forums. Values are presented separately for comments left in banned and non-banned forums, average values for treated and control groups are provided as a reference point. Incidence of slur use is measured as the average number of unique words of a given category per comment for a user. All values are multiplied by 1000 for ease of presentation and to provide a clearer reference to the estimate magnitudes in the main text. The Figure illustrates a) high correlation between exposure to slur type within banned forums and individual slur use and b) significant variation in slur use between banned and non-banned forums when individual exposure is high and relatively small variation when individual exposure is low.

B Additional results on displacement and migration

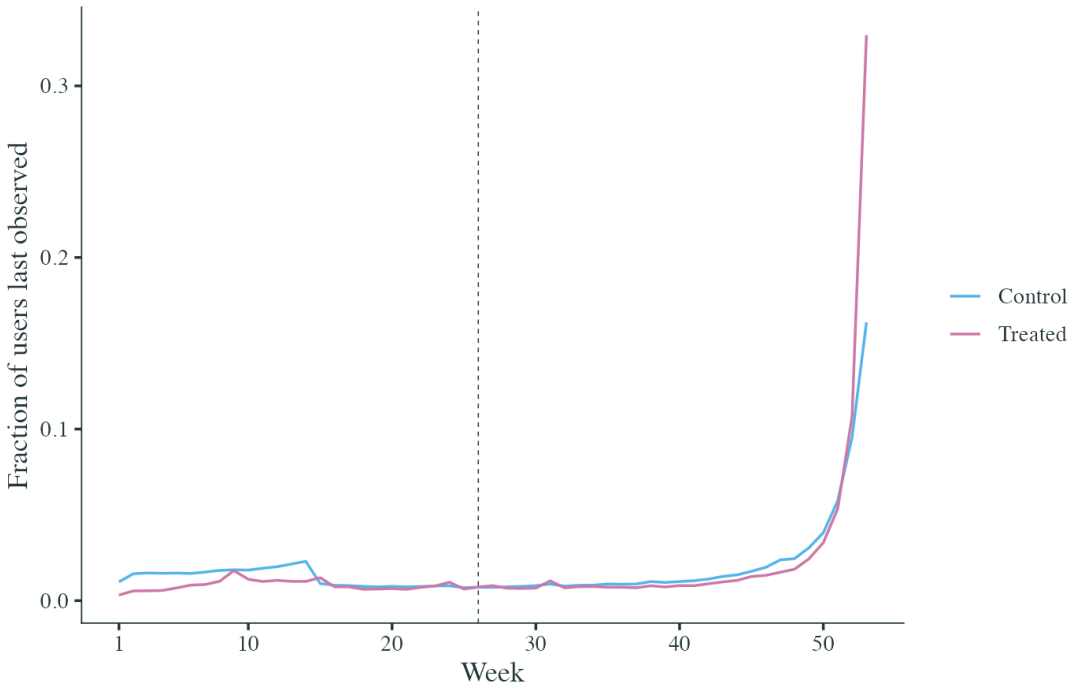
Raw data

Figure B.1: Average weekly number of comments for *core* users vs. control: raw data



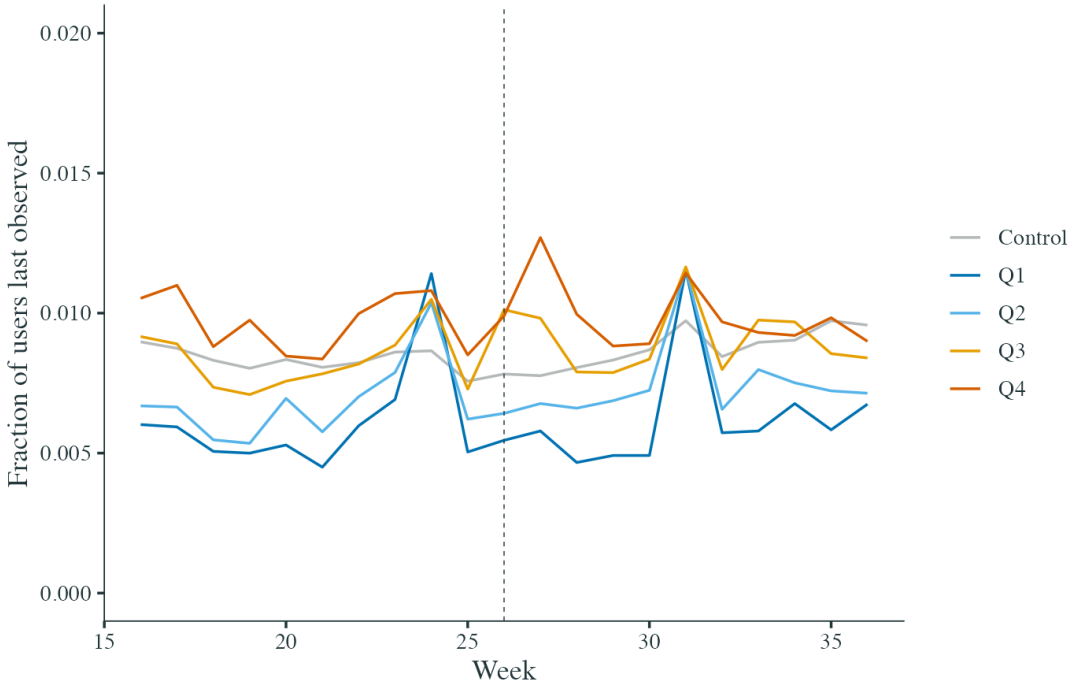
Notes: Figure B.1 plots average weekly number of comments left by most active (*core*) users of banned subreddits (in red) and untreated users (in blue) across Reddit, excluding comments left in banned forums. To ensure comparability, the sample is restricted to users who comment at least 10 times between January and March 2020. The dashed line represents the week of the ban. For clarity of presentation, the values for the untreated group are shifted upwards to match the values for the treated in the week before the policy change. The Figure shows large and sustained increase in the number of comments left by *core* treated users after the policy change.

Figure B.2: Weekly drop-off: treated vs non-treated users



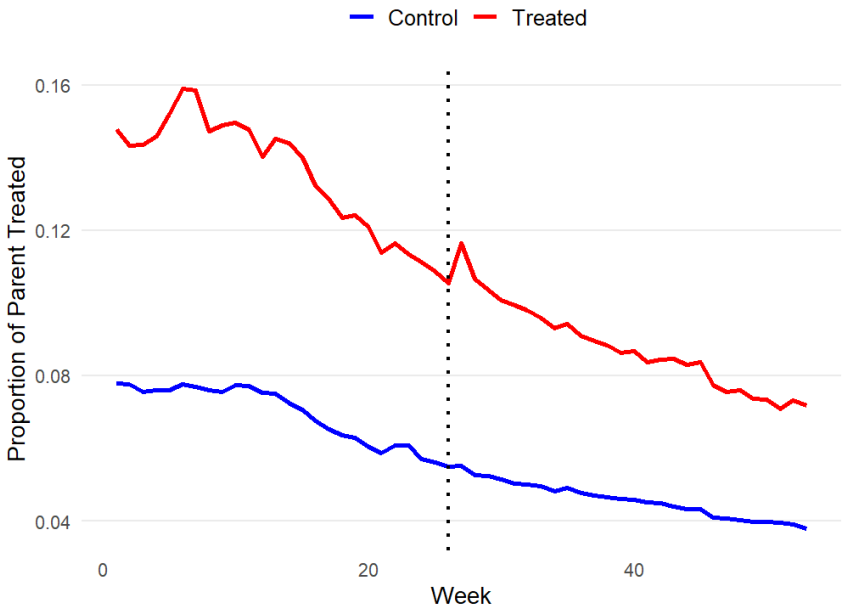
Notes: Figure B.2 shows the fraction of users in treated and non-treated groups (y-axis) last observed in the corresponding week of the year (x-axis). The week before the policy change (*week 26*) is indicated in dashed vertical line. The large uptick towards the end of the sample is mechanical, since the data ends at week 53, so every user is last observed in week 53 at the latest. If it is the case that the results are driven by most extreme treated users simply leaving Reddit after the policy change, we would expect higher drop-off after week 27 – this does not appear to be the case.

Figure B.3: Weekly drop-off: treated vs non-treated users by quartile of ban exposure



Notes: Figure B.3 shows the fraction of users in treated and non-treated groups (y-axis) last observed in the corresponding week of the year (x-axis). The Figure zooms in on the period of 10 weeks before and after the policy change. The week before the policy change (*week 26*) is indicated in dashed vertical line. Treated users are split into quartiles of exposure to the ban ($ShareInBan_i$). The figure confirms that we do not observe significantly higher level of platform exit for core users of banned forums. Higher baseline levels of drop-off for higher quartiles of policy exposure are driven by disproportionate selection of users with lower baseline level of activity in higher quartiles of exposure.

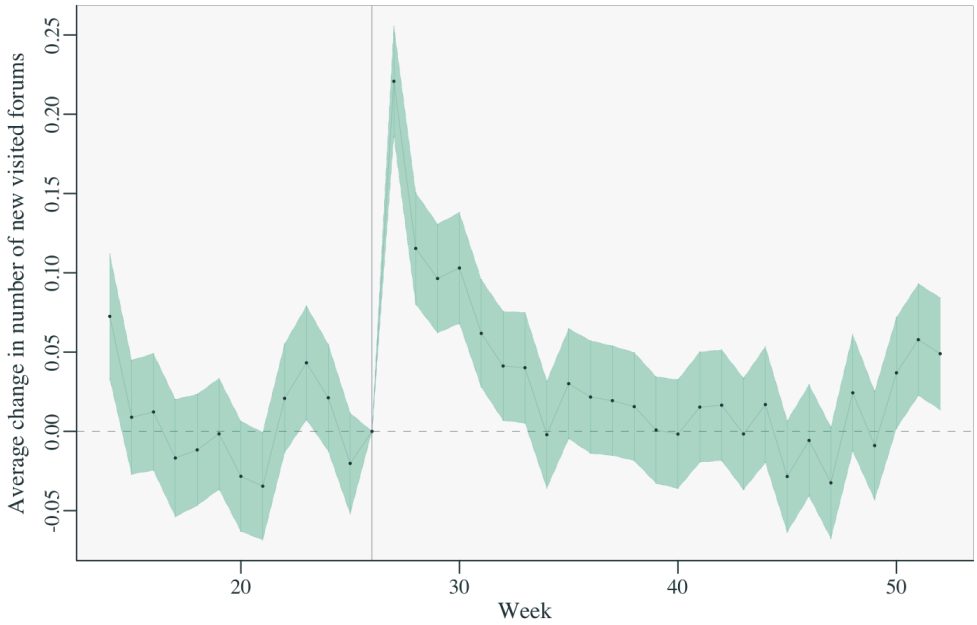
Figure B.4: Probability to reply to a treated user: raw data



Notes: Figure B.4 shows the proportion of comments in each week of the year that are replies to comments left by treated users. This takes the full sample of comments and aggregates them weekly. The continuous decay is due to normal attrition of treated users from the platform.

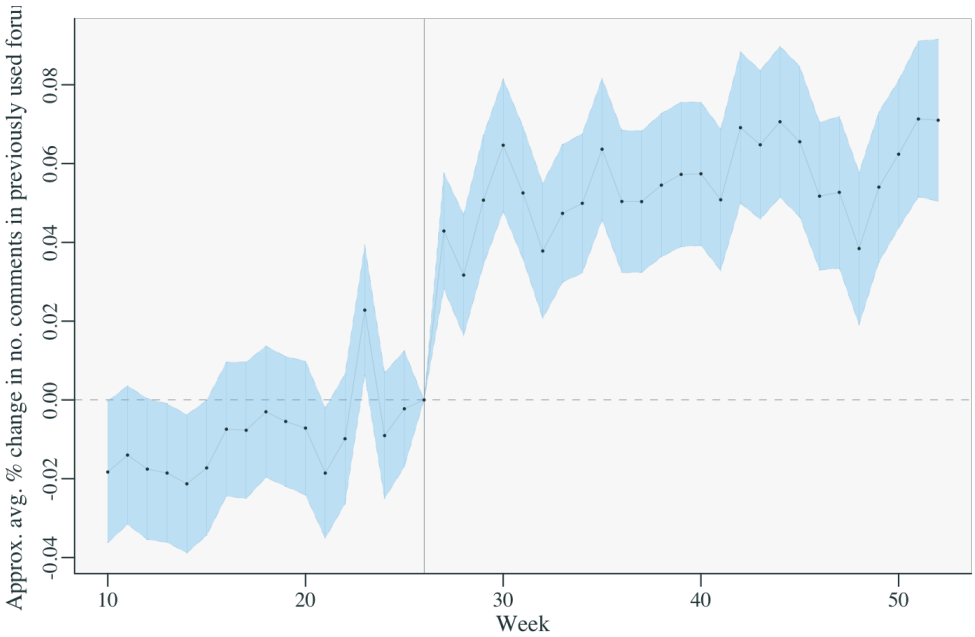
Migration

Figure B.5: Effect of the policy change on the number of new forums visited by *core* treated users: difference-in-differences estimates



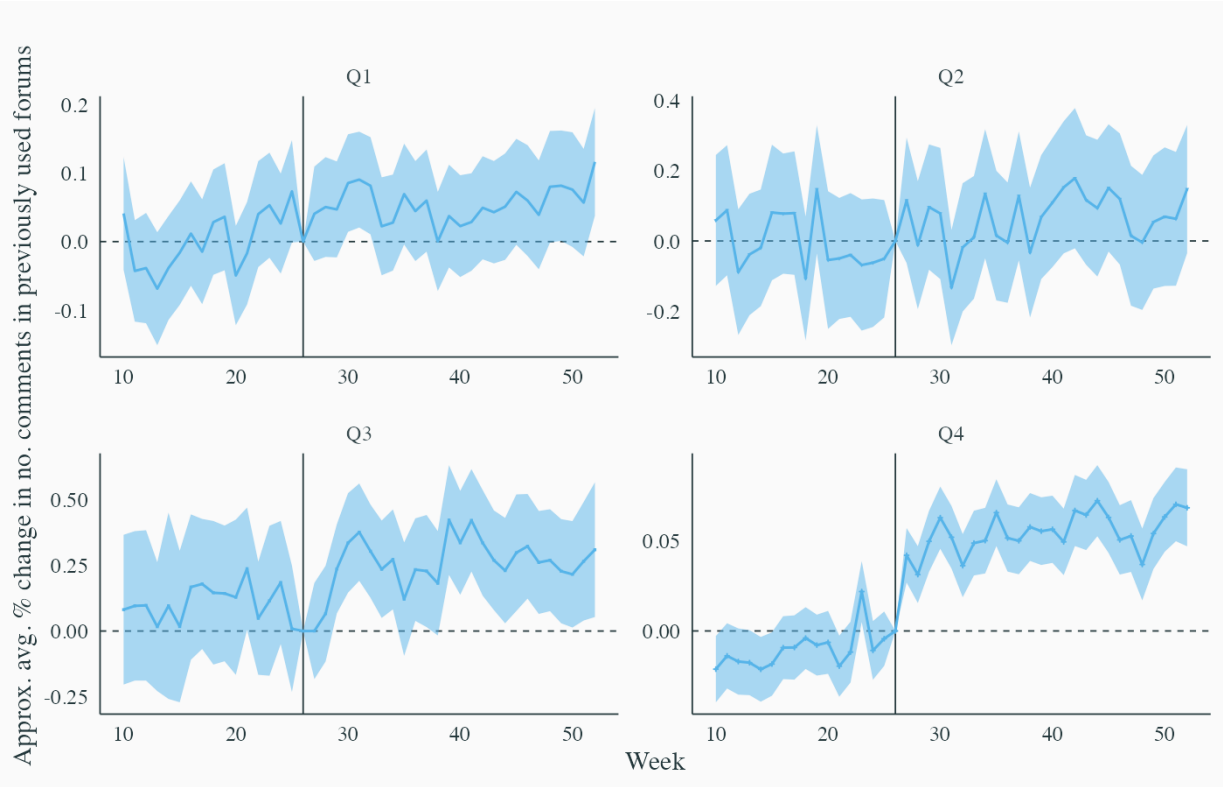
Notes: Figure B.5 shows the event study of the effect of the policy change on the number of new forums visited by *core* treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$. The y-axis represents the average change in number of new forums visited by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level, N=232,126,419.

Figure B.6: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users: difference-in-differences estimates



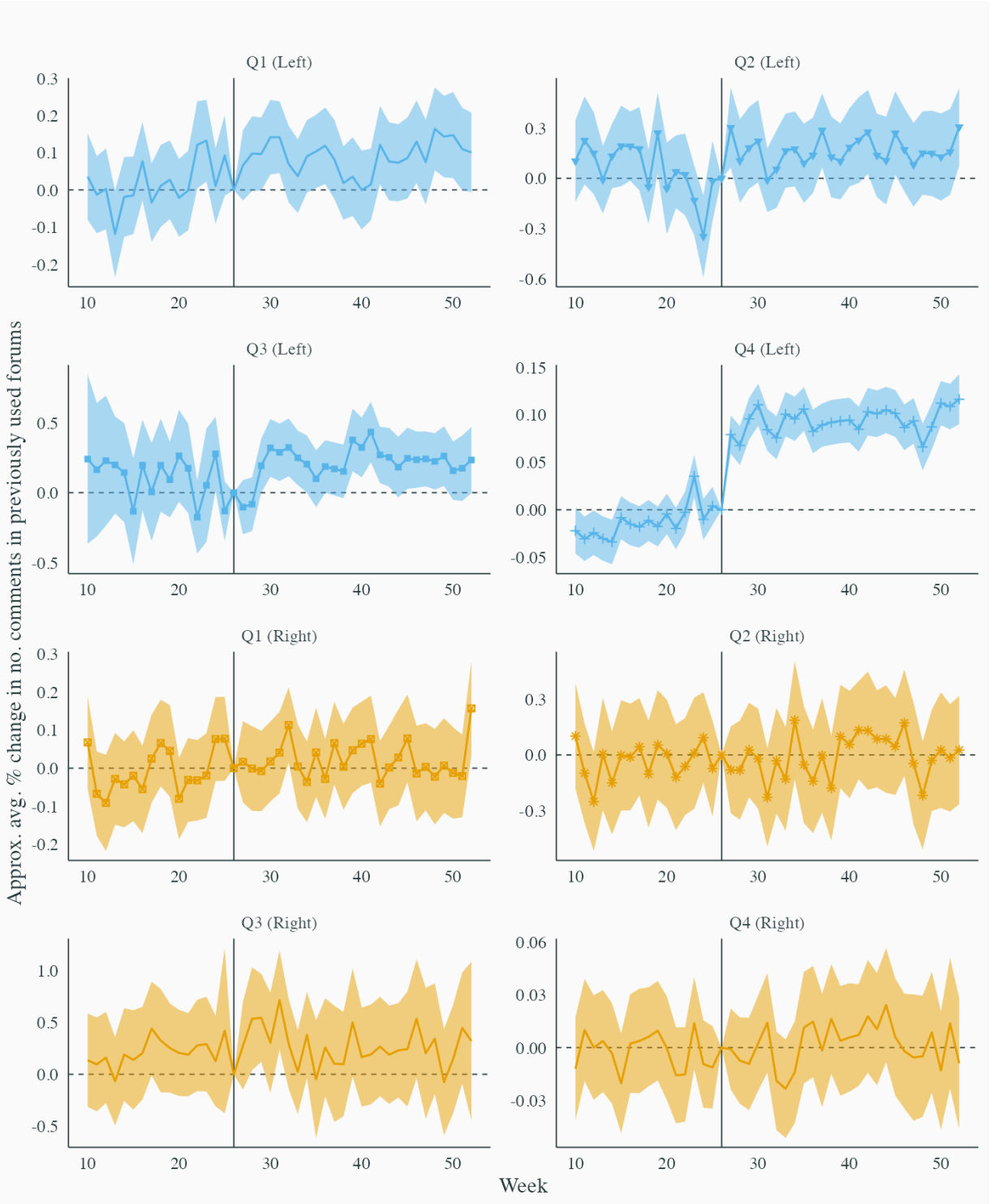
Notes: Figure B.6 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level, N=603,184,977.

Figure B.7: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by quartile of forum hate speech use: difference-in-differences estimates



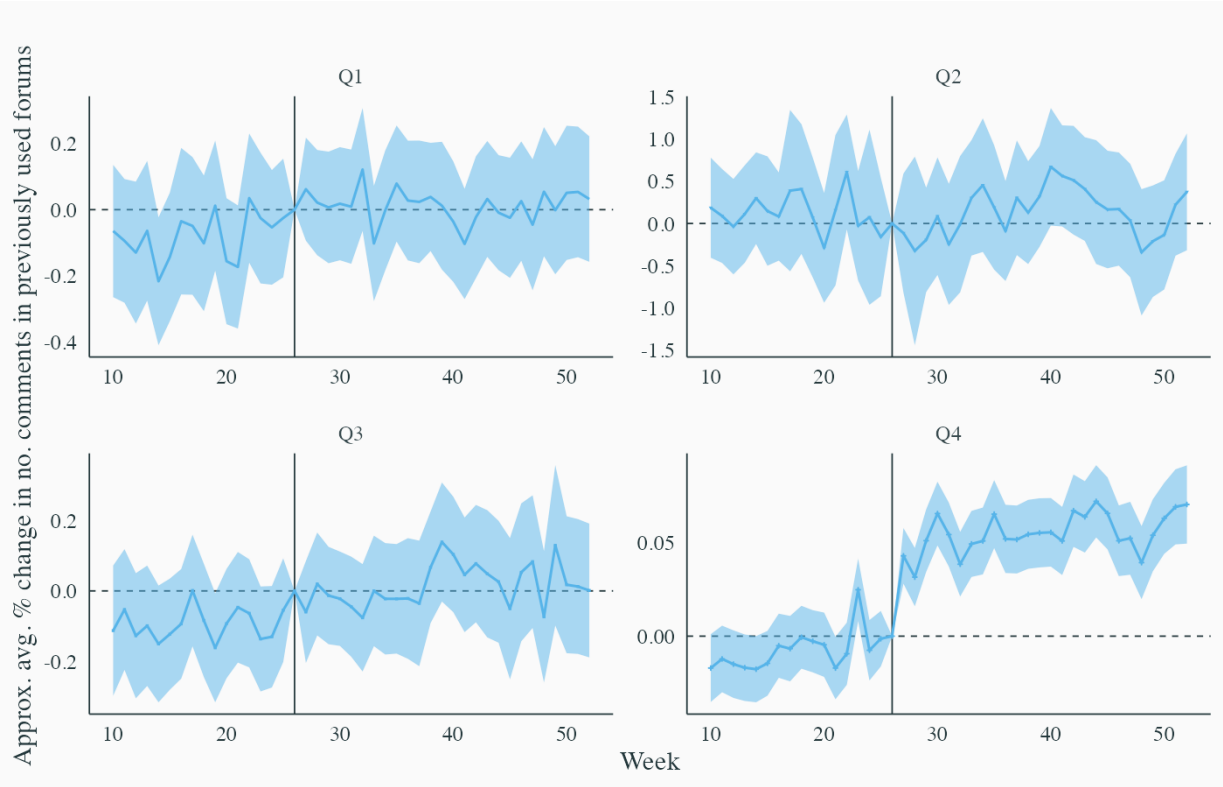
Notes: Figure B.7 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by quartile of forum hate speech use outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for quartiles of forum hate speech use. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.8: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of forum hate speech use: difference-in-differences estimates



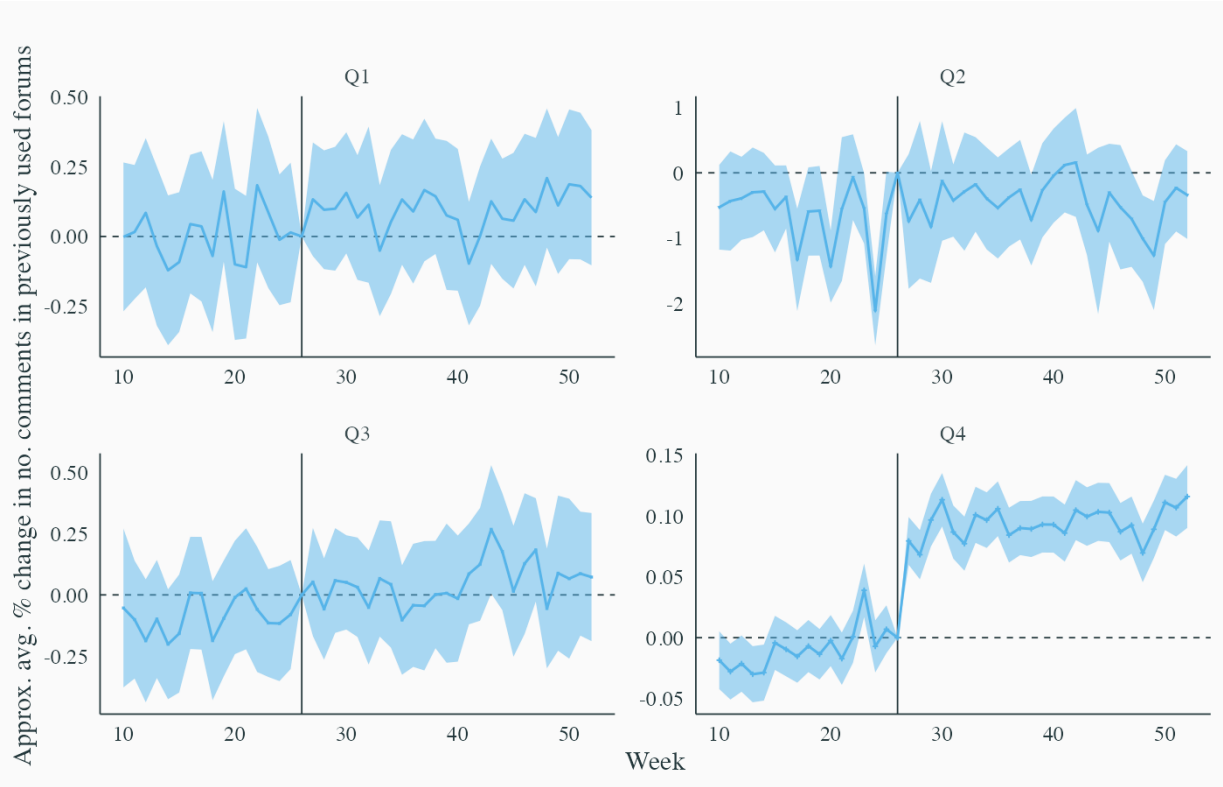
Notes: Figure B.8 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of forum hate speech use outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users and by quartiles of forum hate speech use. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.9: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by quartile of forum toxicity: difference-in-differences estimates



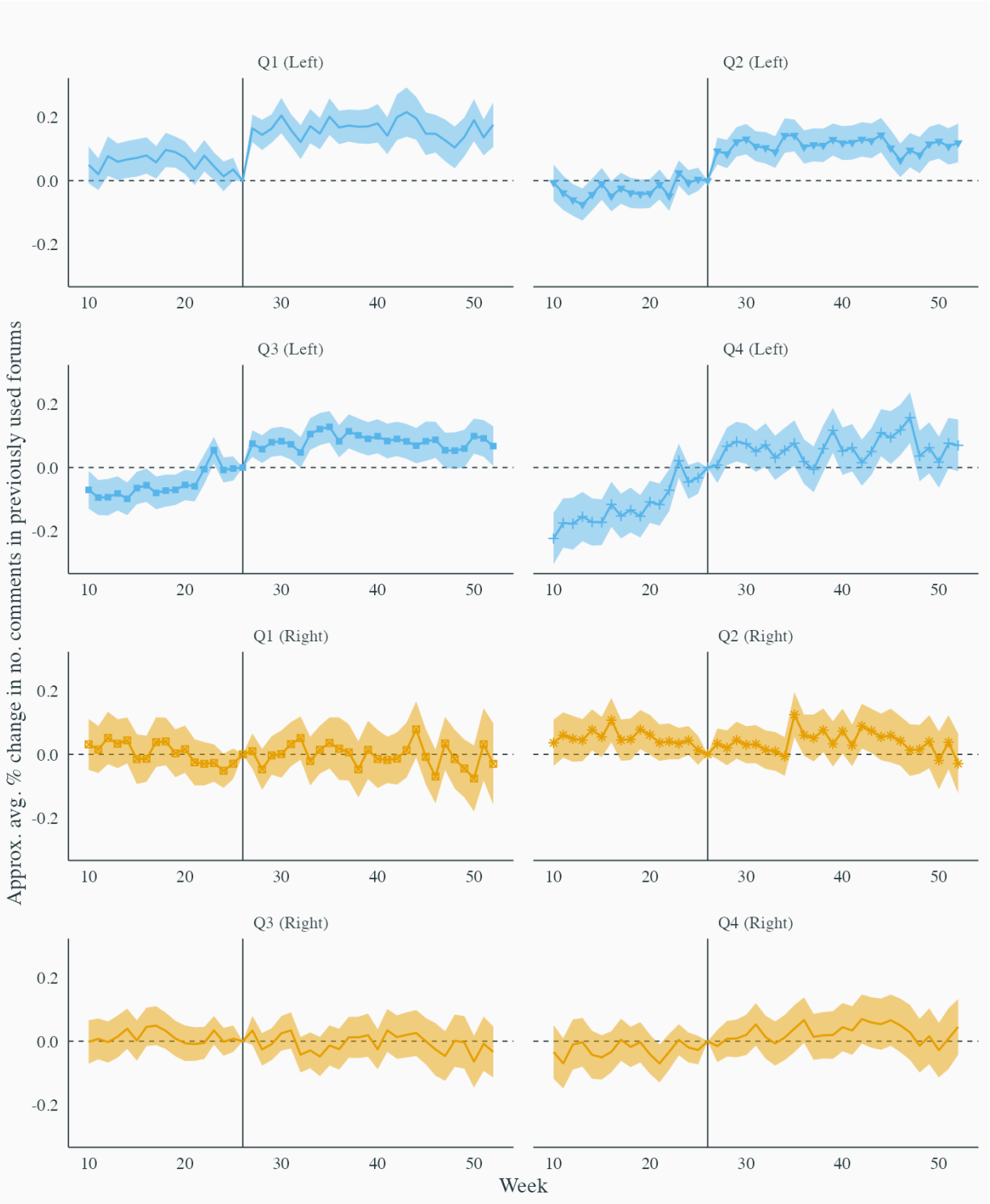
Notes: Figure B.9 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by quartile of forum toxicity outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for quartiles of forum toxicity. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.10: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of forum toxicity: difference-in-differences estimates



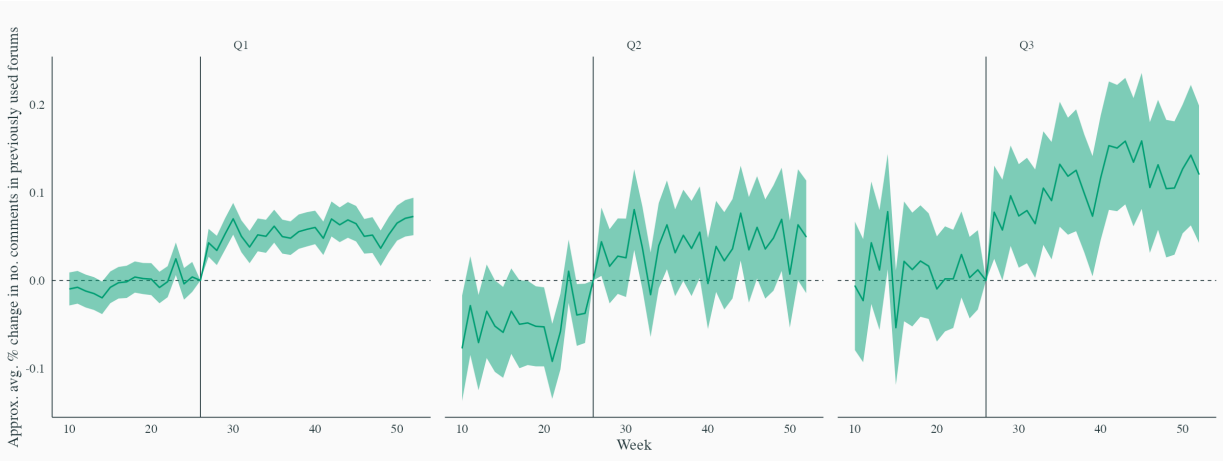
Notes: Figure B.10 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of forum toxicity outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users and by quartiles of forum toxicity. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.11: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of previous use of a forum: difference-in-differences estimates



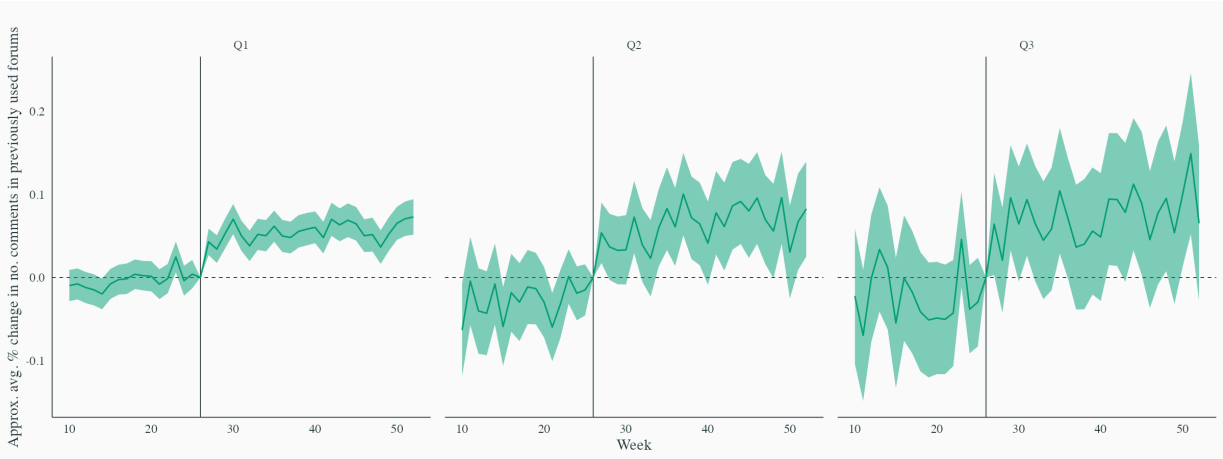
Notes: Figure B.11 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated left- and right-wing users by quartile of previous use of a forum outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users and by quartile of the degree of previous use of a forum. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.12: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (score): difference-in-differences estimates



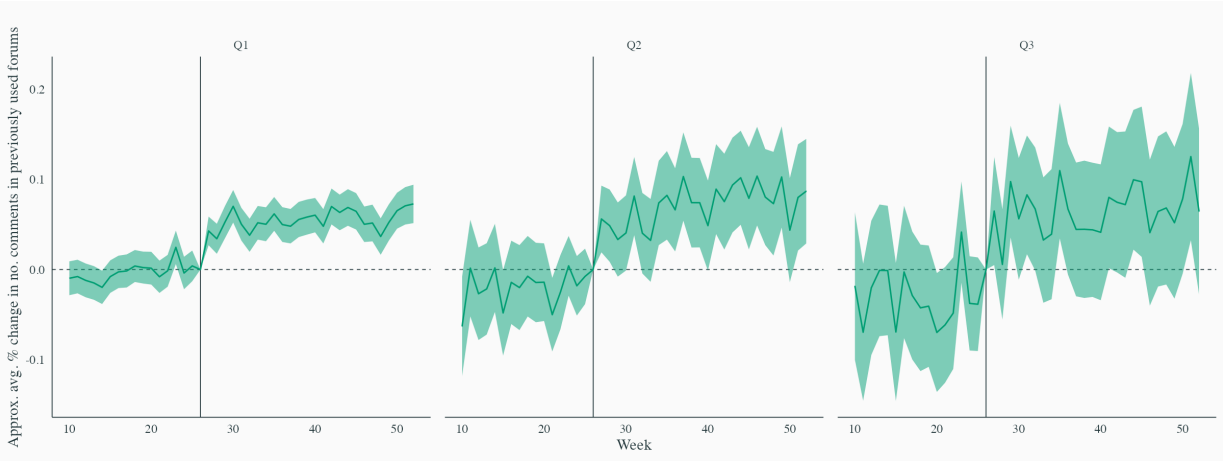
Notes: Figure B.12 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (score) outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately depending on level of engagement user received in a forum prior to the policy change (measured by the average comment score). The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.13: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (direct replies): difference-in-differences estimates



Notes: Figure B.13 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (direct replies) outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately depending on level of engagement user received in a forum prior to the policy change (measured by the average number of direct replies). The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

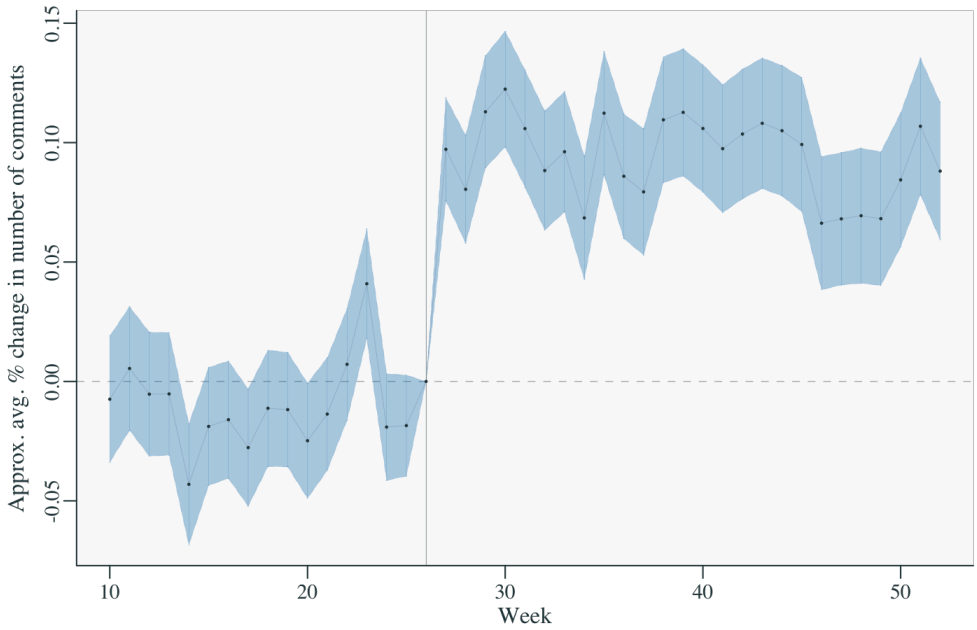
Figure B.14: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (total replies): difference-in-differences estimates



Notes: Figure B.14 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users by user's engagement in a forum (total replies) outside banned forums. Coefficients are produced by estimating Equation 4 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately depending on level of engagement user received in a forum prior to the policy change (measured by the average number of total replies). The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

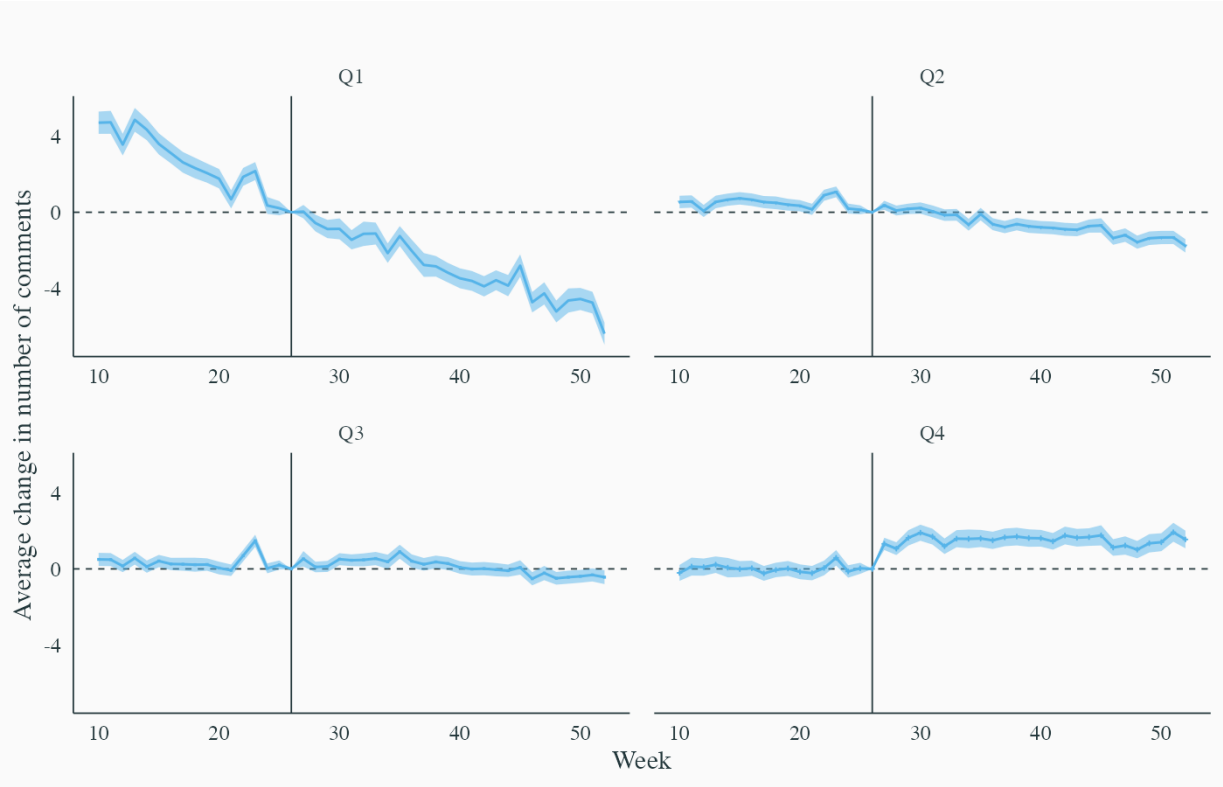
Alternative specifications & user heterogeneity

Figure B.15: Effect of the policy change on the log of number of comments left by *core* treated users: difference-in-differences estimates



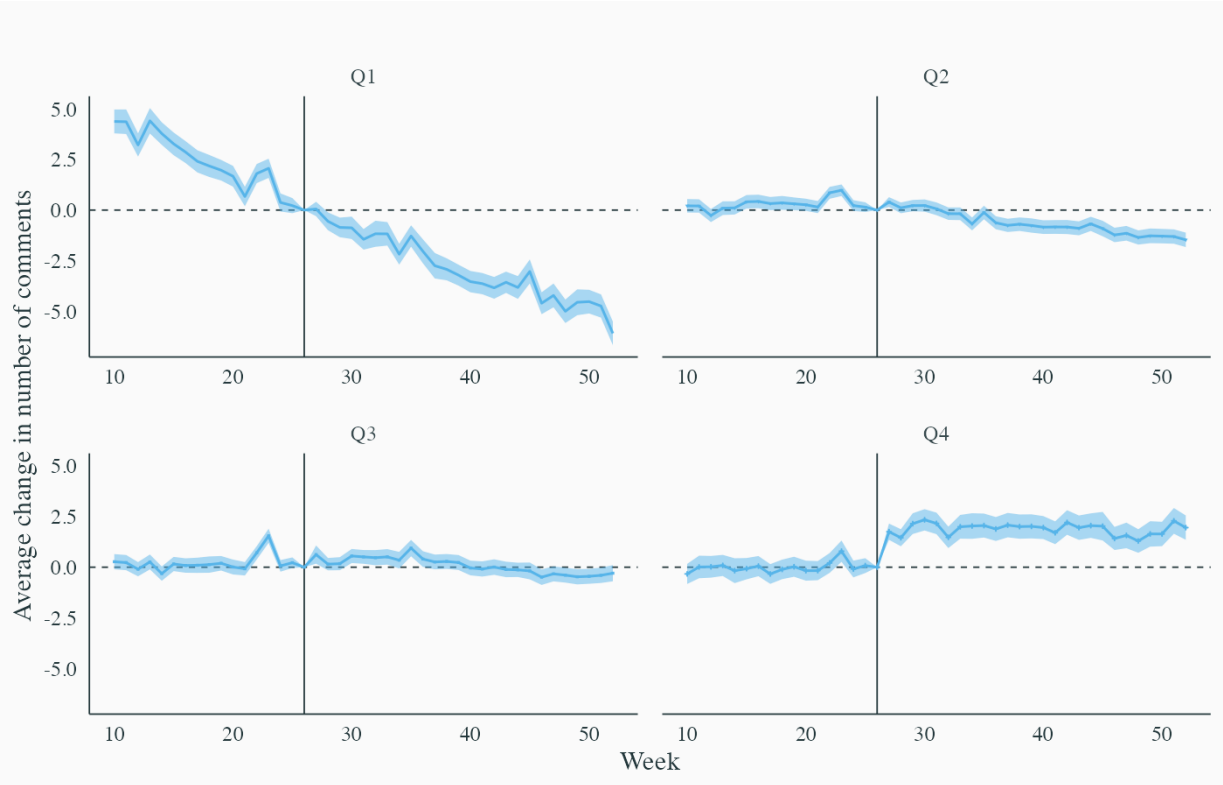
Notes: Figure B.15 shows the event study of the effect of the policy change on the log of number of comments left by *core* treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$. The y-axis represents the approximate average percentage change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level, N=232,126,419. Equivalent results using raw number of comments as the outcome variable are presented in Figure 1.

Figure B.16: Effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban: difference-in-differences estimates



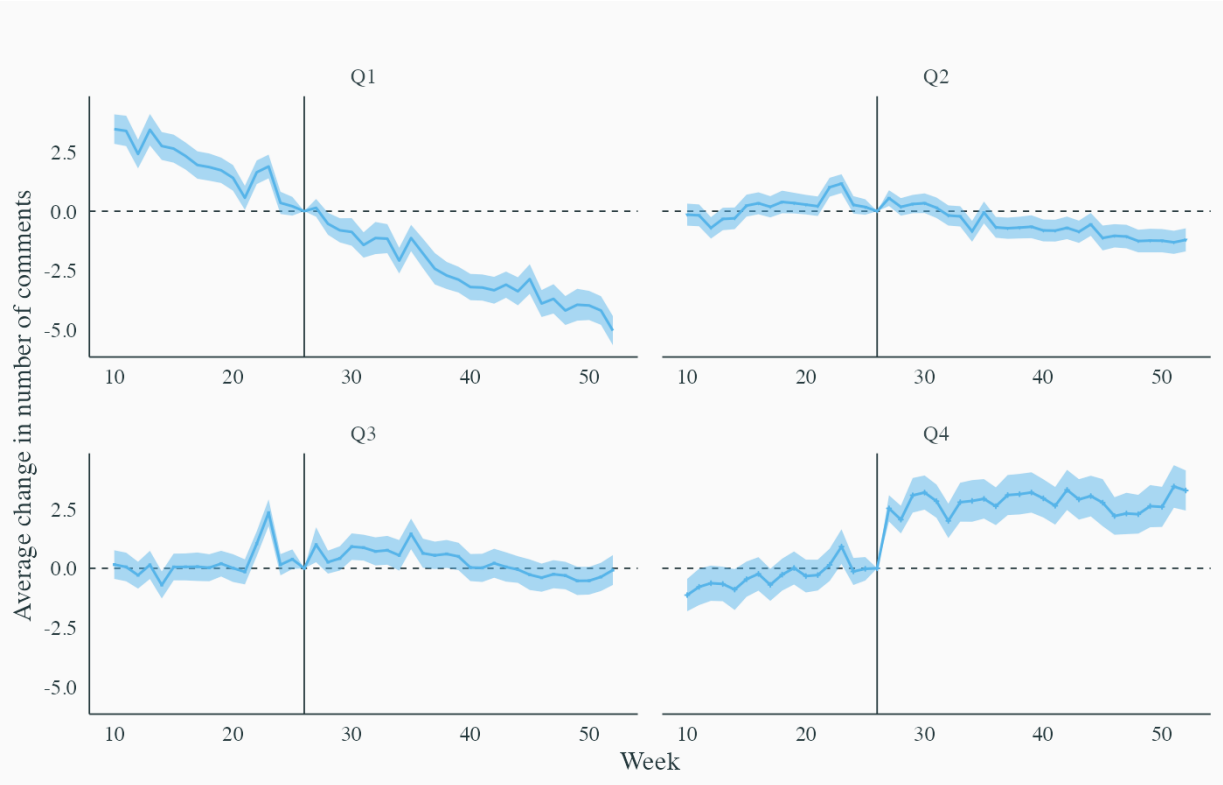
Notes: Figure B.16 shows the event study of the effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban outside banned forums. Coefficients are produced by estimating Equation 1 separately by quartile of $ShareInBan_i$. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using log of number of comments as the outcome variable are presented in Figure B.19. Results from a triple difference specification are shown in Figure B.20.

Figure B.17: Effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban: difference-in-differences estimates (10 comment cutoff)



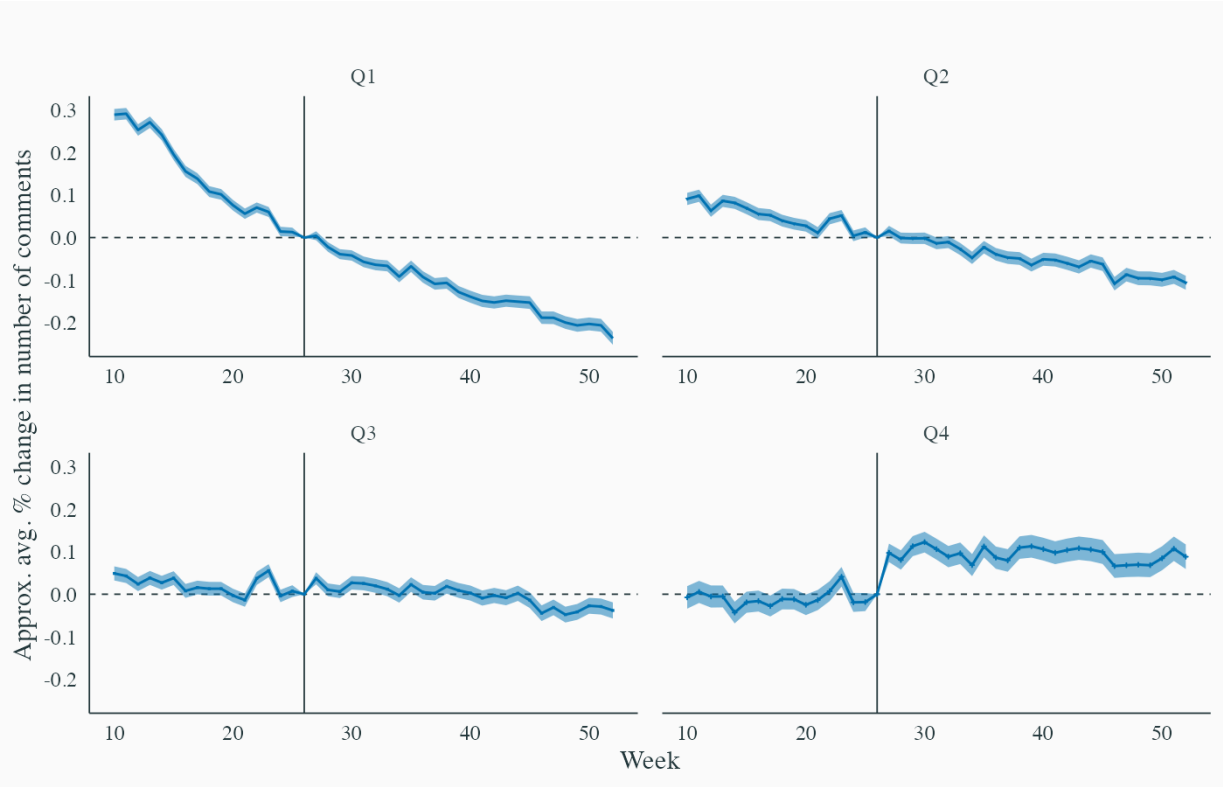
Notes: Figure B.17 shows the event study of the effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban outside banned forums. Coefficients are produced by estimating Equation 1 separately by quartile of $ShareInBan_i$. Sample is restricted to users who left at least 10 comments between January and March 2020 (see Section Y for discussion). The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.18: Effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban: difference-in-differences estimates (50 comment cutoff)



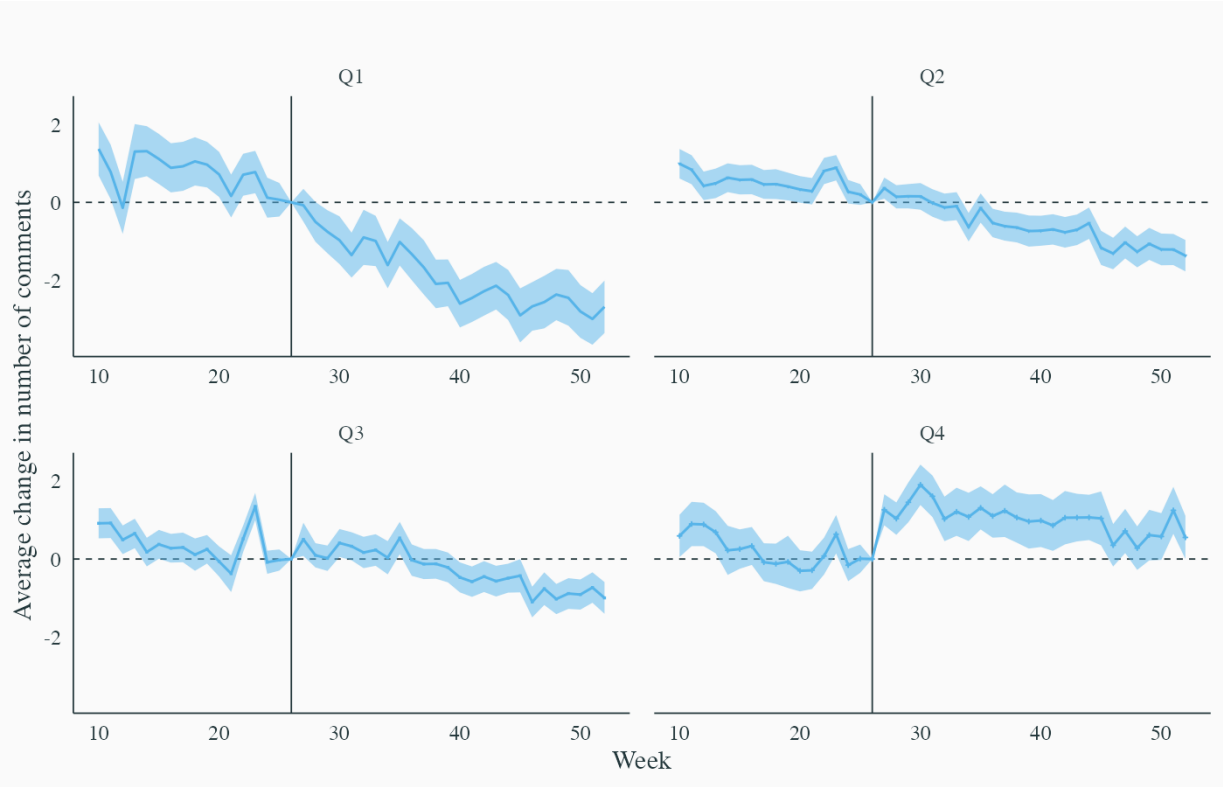
Notes: Figure B.18 shows the event study of the effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban outside banned forums. Coefficients are produced by estimating Equation 1 separately by quartile of $ShareInBan_i$. Sample is restricted to users who left at least 50 comments between January and March 2020 (see Section Y for discussion). The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.19: Effect of the policy change on the log of number of comments left by treated users depending on the degree of exposure to the ban: difference-in-differences estimates



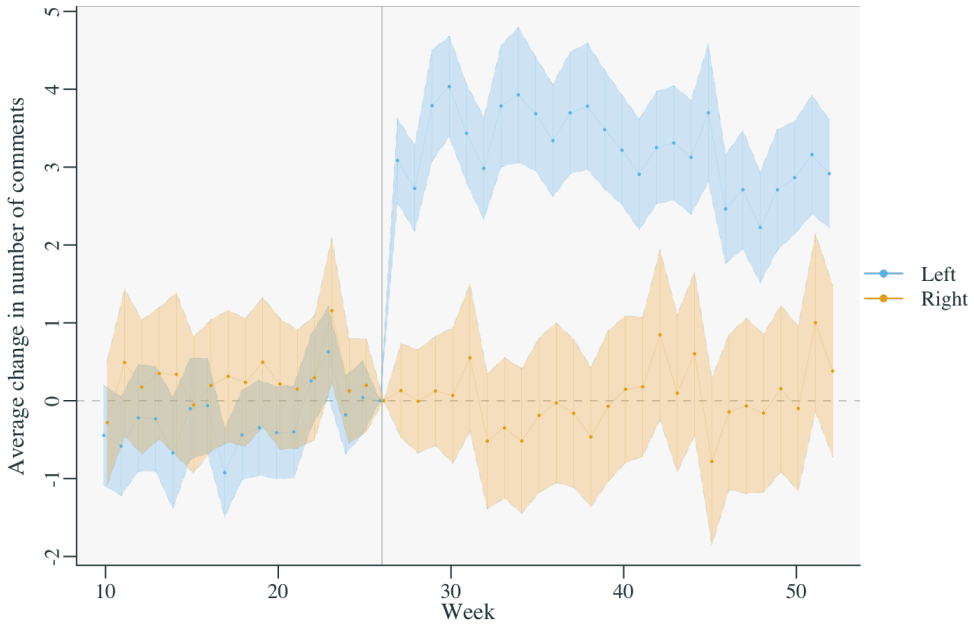
Notes: Figure B.19 shows the event study of the effect of the policy change on the log of number of comments left by treated users depending on the degree of exposure to the ban outside banned forums. Coefficients are produced by estimating Equation 1 separately by quartile of $ShareInBan_i$. The y-axis represents the approximate average percentage change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using raw number of comments as the outcome variable are presented in Figure B.16.

Figure B.20: Effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban: triple difference estimates



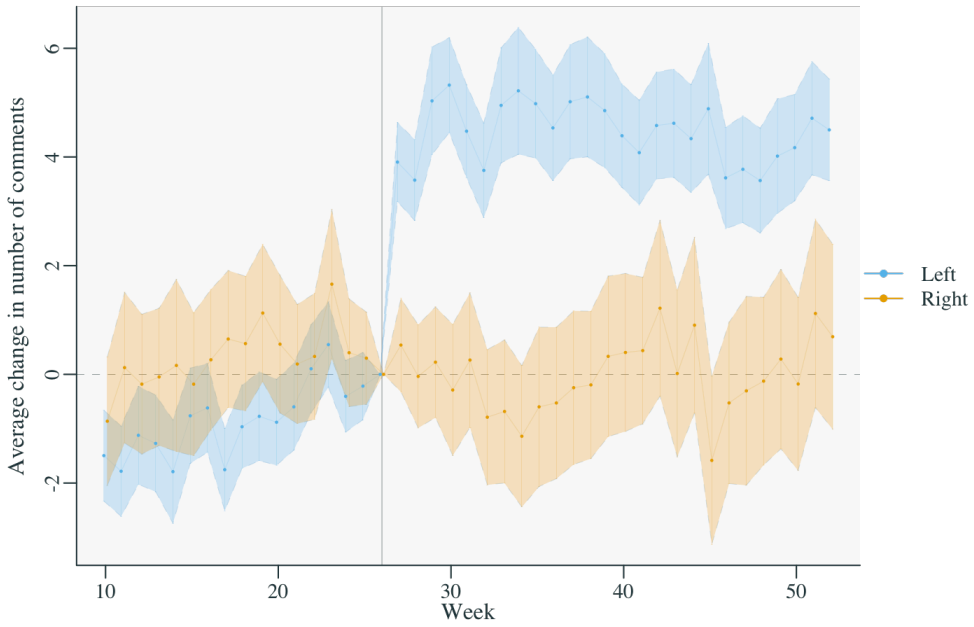
Notes: Figure B.20 shows the event study of the effect of the policy change on the number of comments left by treated users depending on the degree of exposure to the ban outside banned forums. Coefficients are produced by estimating triple-difference version of Equation 1 separately by quartile of $ShareInBan_i$. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Results from a difference-in-differences specification are shown in Figure B.16.

Figure B.21: Effect of the policy change on the number of comments left by *core* left- and right-wing treated users: difference-in-differences estimates (10 comment cutoff)



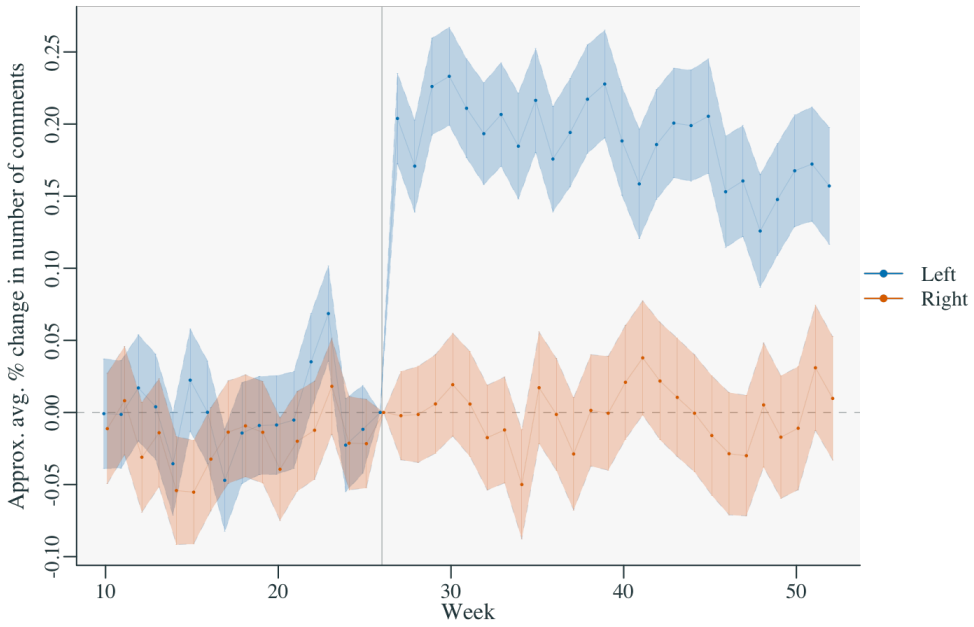
Notes: Figure B.21 shows the event study of the effect of the policy change on the number of comments left by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. Sample is restricted to users who left at least 10 comments between January and March 2020 (see Section Y for discussion). The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.22: Effect of the policy change on the number of comments left by *core* left- and right-wing treated users: difference-in-differences estimates (50 comment cutoff)



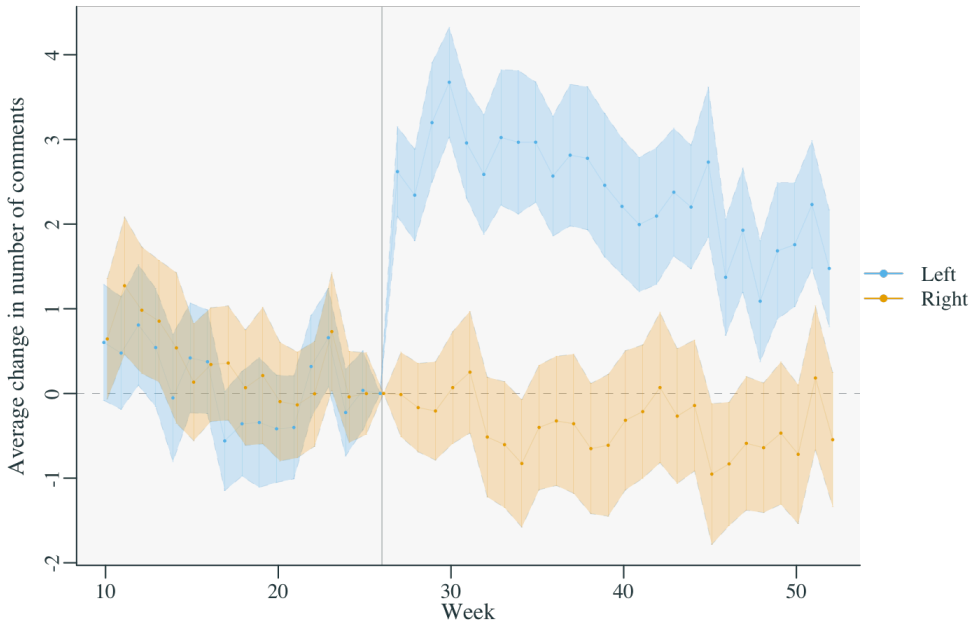
Notes: Figure B.22 shows the event study of the effect of the policy change on the number of comments left by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. Sample is restricted to users who left at least 50 comments between January and March 2020 (see Section Y for discussion). The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.23: Effect of the policy change on the log of number of comments left by *core* left- and right-wing treated users: difference-in-differences estimates



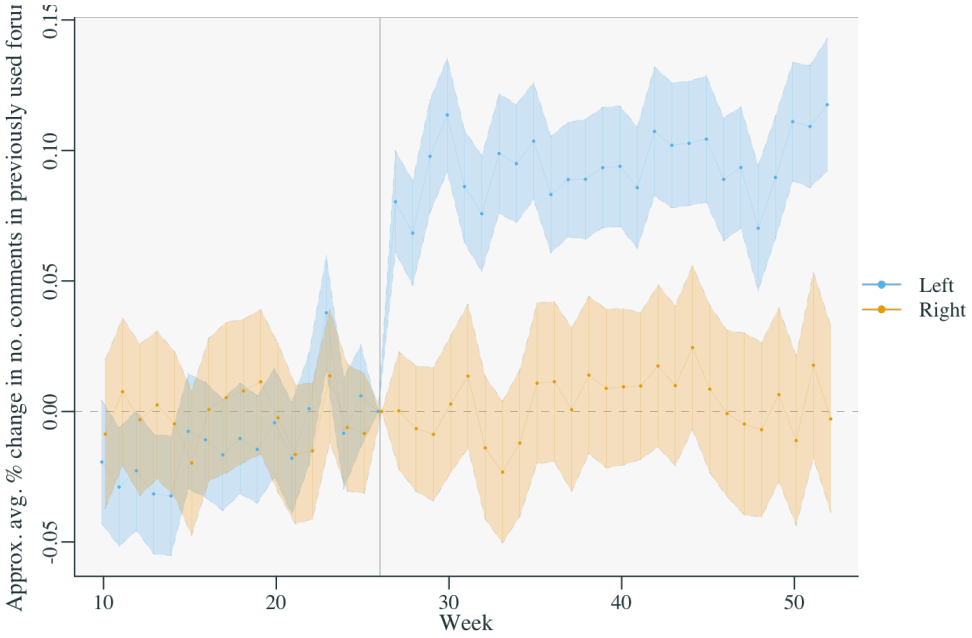
Notes: Figure B.23 shows the event study of the effect of the policy change on the log of number of comments left by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the approximate average percentage change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using raw number of comments as the outcome variable are presented in Figure 2.

Figure B.24: Effect of the policy change on the number of comments left by *core* left- and right-wing treated users: triple difference estimates



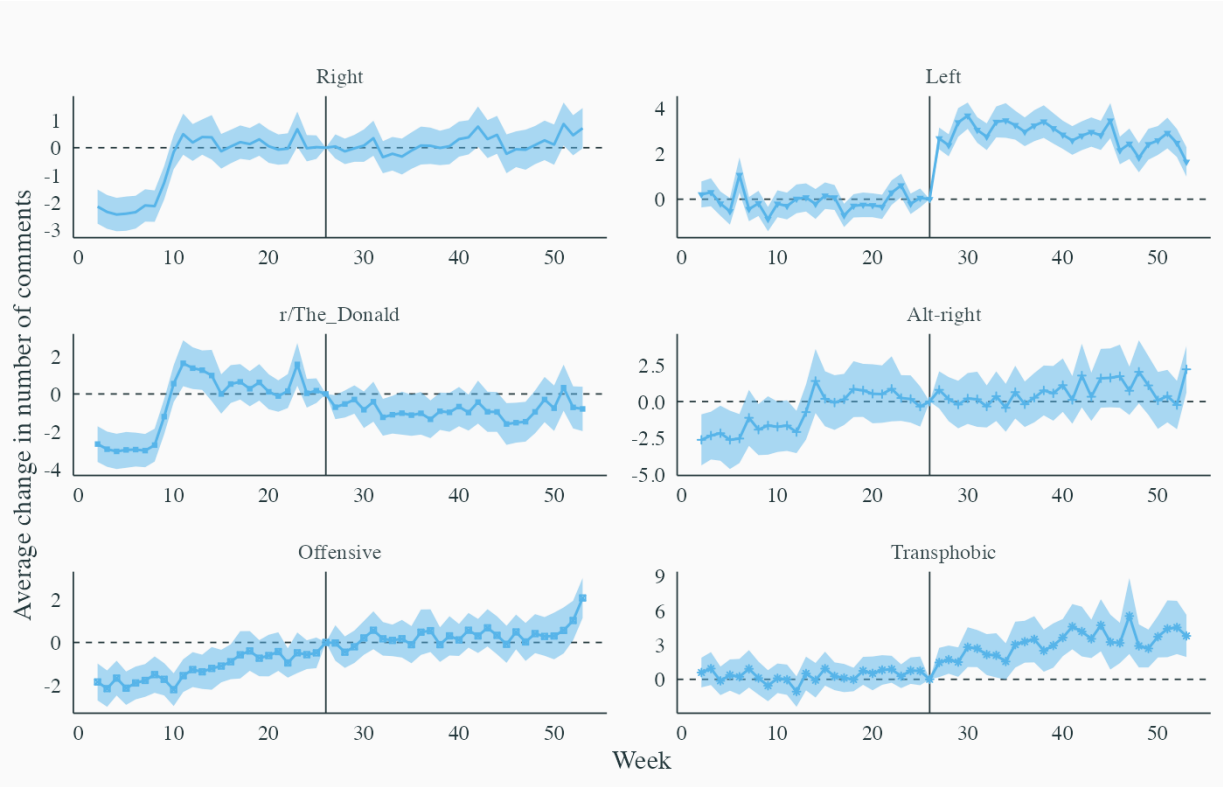
Notes: Figure B.24 shows the event study of the effect of the policy change on the number of comments left by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating triple-difference version of Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Results from a difference-in-differences specification are shown in Figure 2.

Figure B.25: Effect of the policy change on the log of number of comments left in previously used forums by *core* left- and right-wing treated users: difference-in-differences estimates



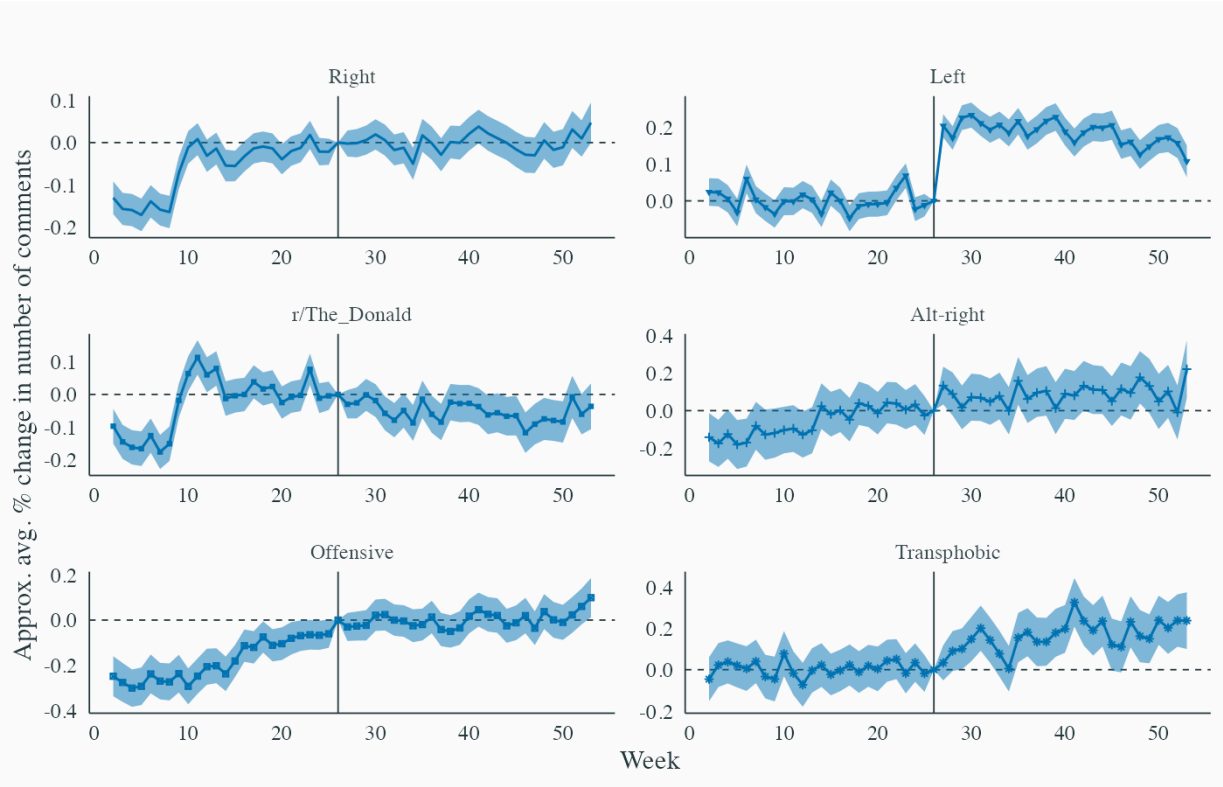
Notes: Figure B.25 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* left- and right-wing treated users outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for left- and right-wing users. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.26: Effect of the policy change on the number of comments left by *core* treated users exposed to specific types of banned forums: difference-in-differences estimates



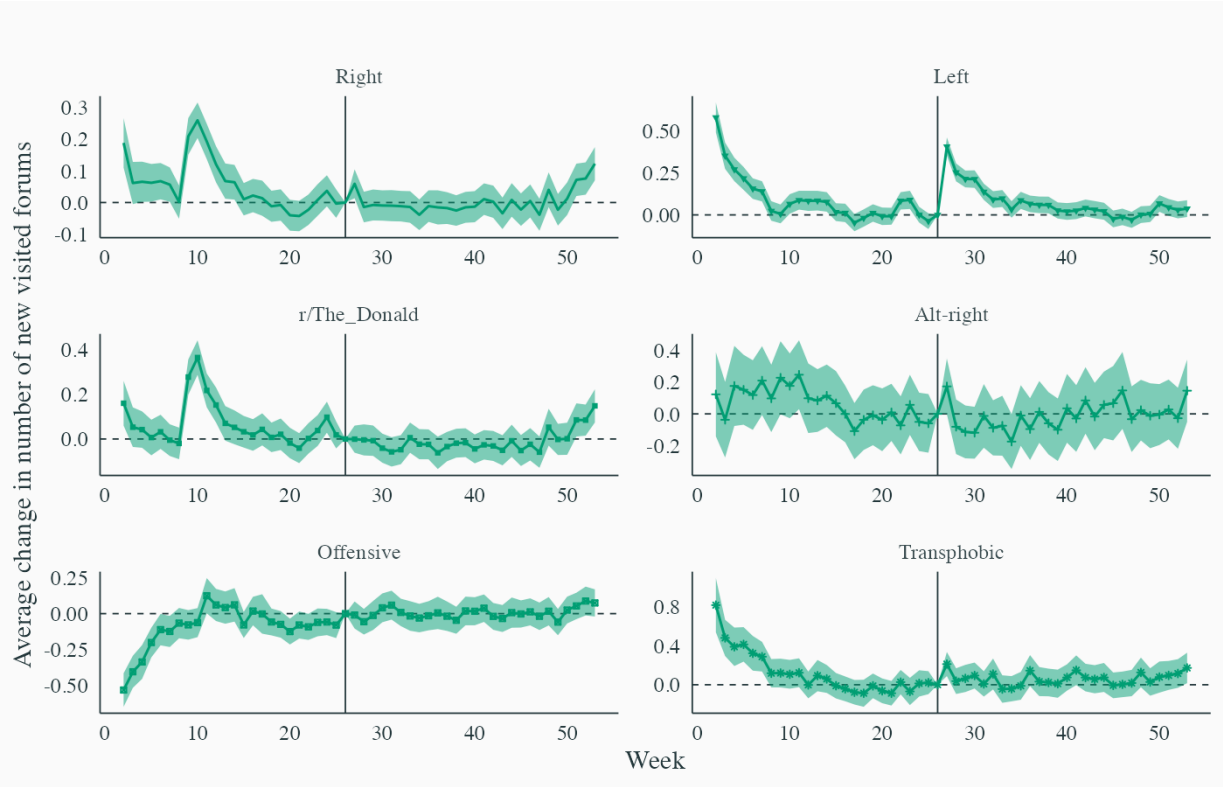
Notes: Figure B.26 shows the event study of the effect of the policy change on the number of comments left by *core* treated users exposed to specific types of banned forums outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for users exposed to different categories of banned forums. The y-axis represents the average change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using log of number of comments as the outcome variable are presented in Figure B.27.

Figure B.27: Effect of the policy change on the log of number of comments left by *core* treated users exposed to specific types of banned forums: difference-in-differences estimates



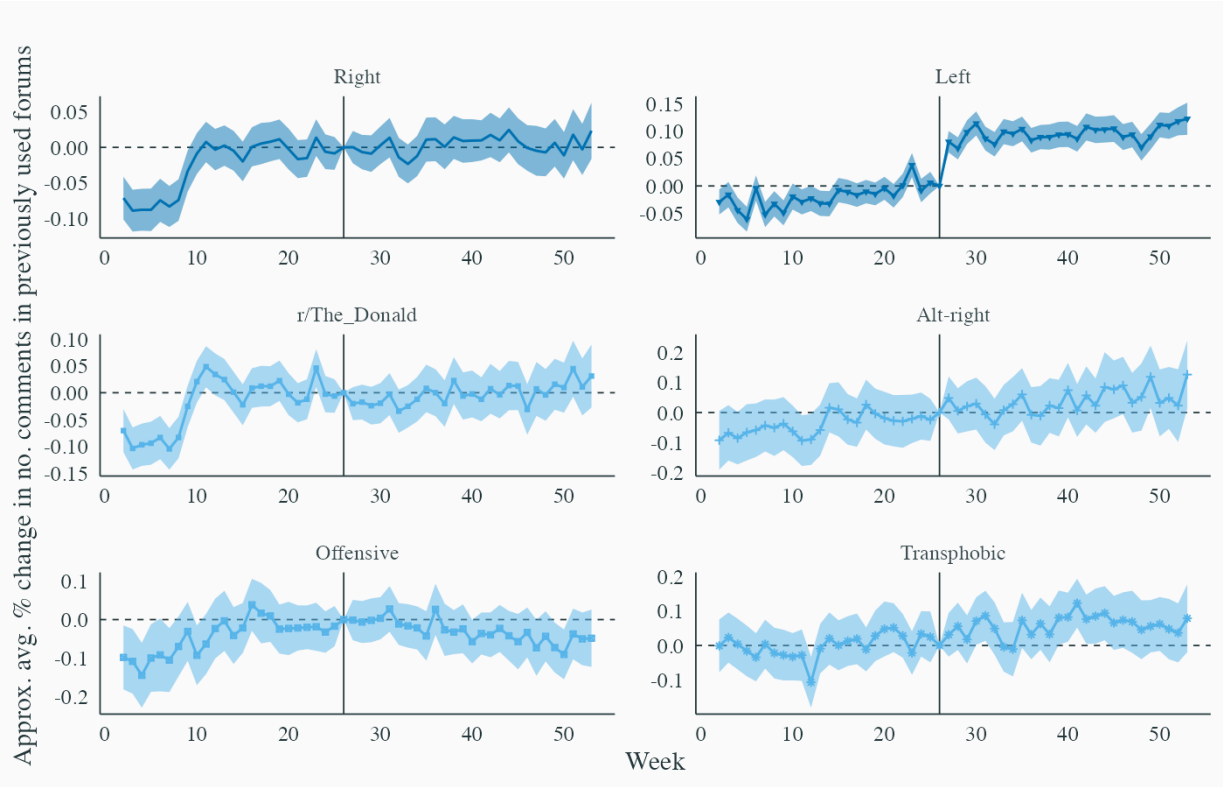
Notes: Figure B.27 shows the event study of the effect of the policy change on the log of number of comments left by *core* treated users exposed to specific types of banned forums outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for users exposed to different categories of banned forums. The y-axis represents the approximate average percentage change in number of comments left by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. Equivalent results using raw number of comments as the outcome variable are presented in Figure B.26.

Figure B.28: Effect of the policy change on the number of new forums visited by *core* treated users exposed to specific types of banned forums: difference-in-differences estimates



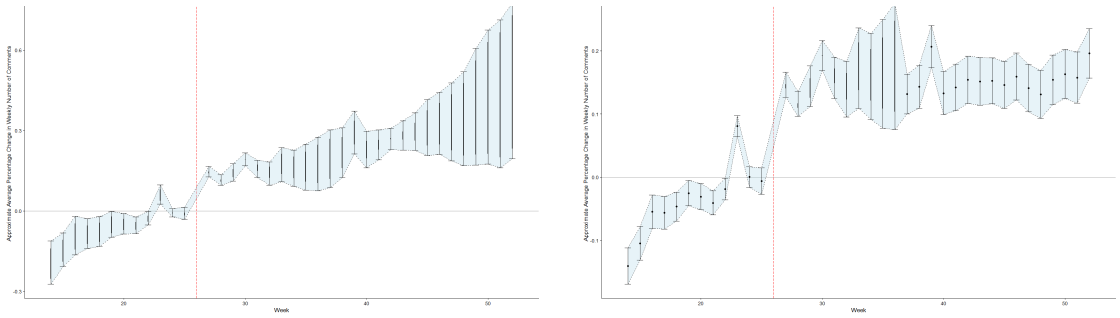
Notes: Figure B.28 shows the event study of the effect of the policy change on the number of new forums visited by *core* treated users exposed to specific types of banned forums outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for users exposed to different categories of banned forums. The y-axis represents the average change in number of new forums visited by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.29: Effect of the policy change on the log of number of comments left in previously used forums by *core* treated users exposed to specific types of banned forums: difference-in-differences estimates



Notes: Figure B.29 shows the event study of the effect of the policy change on the log of number of comments left in previously used forums by *core* treated users exposed to specific types of banned forums outside banned forums. Coefficients are produced by estimating Equation 1 on the subsample of users in 4^{th} quartile of $ShareInBan_i$, separately for users exposed to different categories of banned forums. The y-axis represents the approximate average percentage change in number of comments left within previously used forums by a user; the x-axis denotes weeks. The base category is the week before the policy change (*week 26*) indicated by the vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level.

Figure B.30: Effect on log number of comments: event study with sample restriction, difference-in-differences specification, Lee bound estimates



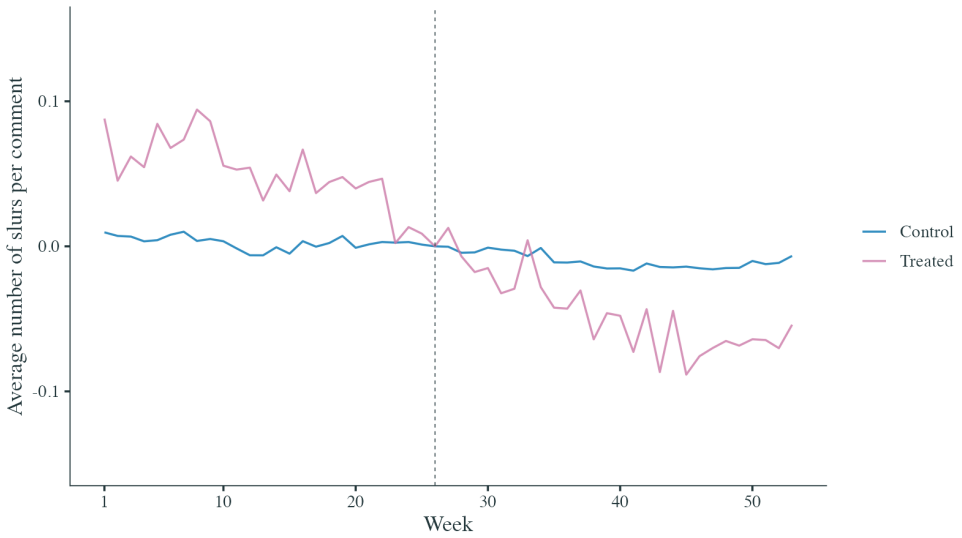
(a) Conservative (b) Baseline

Notes: Figure B.30 presents the results of estimating Equation 1using a Lee bounds style (Lee 2009) trimming procedure. The exercise aims to ensure that the results are not driven by platform exit (although, as we discuss in the paper, the magnitudes of drop-off are small). The drop-off rates for each week are calculated and the top or bottom proportion of either treatment (4th quartile) or control are trimmed to match the samples. Standard errors are calculated by bootstrapping the full procedure. In Panel (a) the event study approach assumes all differential drop-off (even before the ban) could be caused by the policy change and could affect the intensive margin. This is the reason the identified set grows bigger across weeks, as there is higher relative drop-off over time. Panel (b) restricts the potential for set identification to subsample after the policy change and weeks where the treated users have a higher drop-off than the control group - that is, assuming that the policy could only affect drop-off of the treated users after the ban. The sample is restricted to users who left at least 50 comments between January and March 2020. The results confirm that the effect is not driven by platform exit and that estimates from Equation 1 reflect the intensive margin effect of the policy.

C Additional results on language change

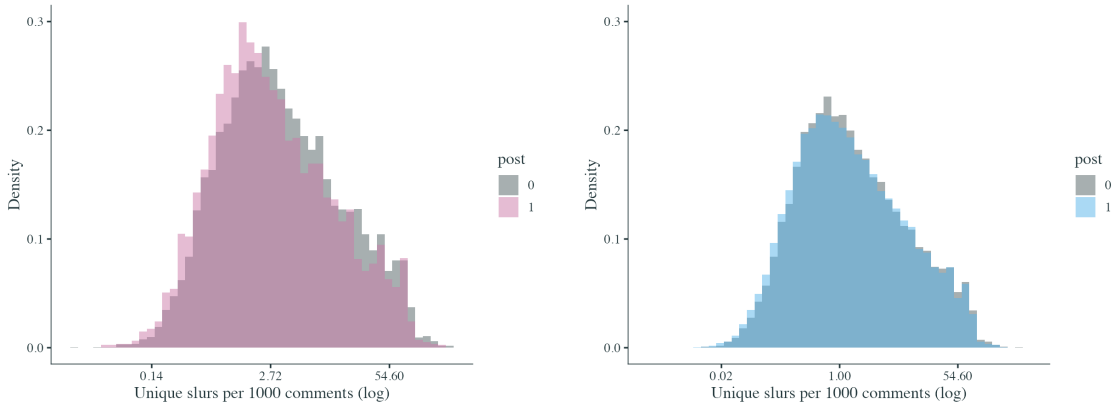
Raw data

Figure C.1: Incidence of hate speech in comments by treated *vs* non-treated users: raw data



Notes: Figure C.1 shows the average number of unique slurs per comment used by members of treated (pink) and control (blue) groups over the weeks of the sample. The values are produced using trimmed ($K = 3$) estimates of slur use (see Section W for discussion). For clarity of presentation the values for both groups are shifted to be equal to zero in the week before the policy change (*week 26*) indicated by dashed vertical line.

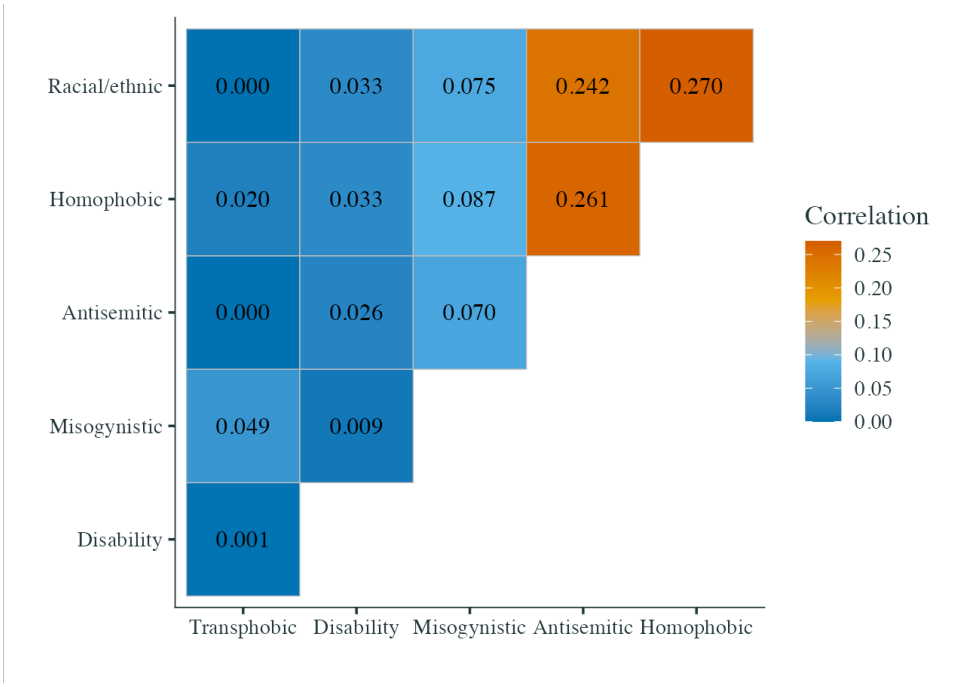
Figure C.2: Average incidence of slur use by treated and control users within a subreddit before and after the policy change



Treated Control

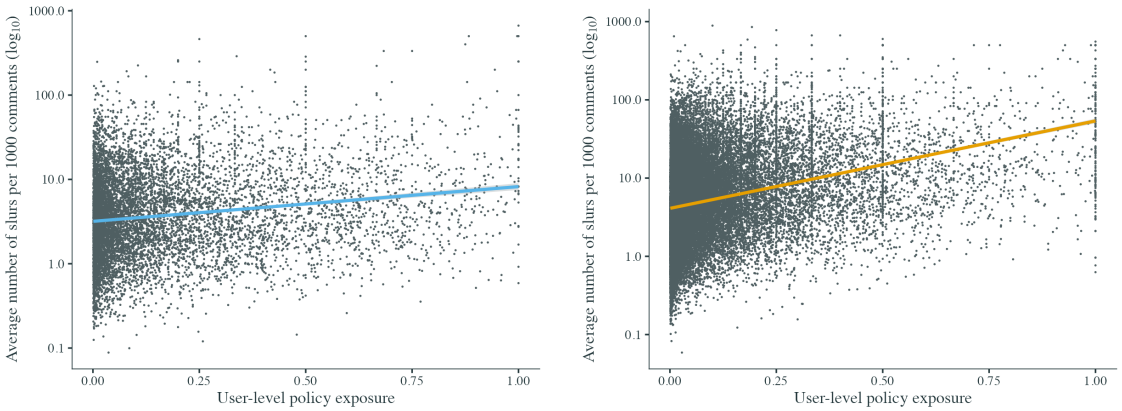
Notes: Figure C.2 illustrates the distribution of average incidence of slur use within a subreddit, separately for users in treated (left panel) and control (right panel) groups. Average is taken separately over weeks before the policy (*post=0*) and weeks after the policy (*post=1*).

Figure C.3: Correlation of slur incidence within forums by category



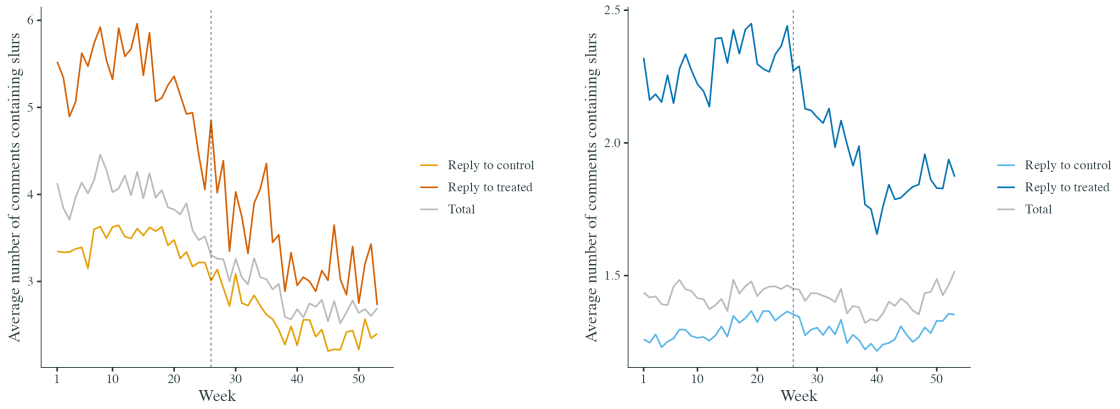
Notes: Figure C.3 displays correlation in of incidence of hate speech across slur category. Together with descriptive statistics on incidence of hate speech by forum from previous section (Table A.3), the data illustrates that the hate speech of different types, with the exception of racial and homophobic slurs, does not tend to be used in the same discussions.

Figure C.4: Correlation between treated users' policy exposure and hate speech use outside banned forums



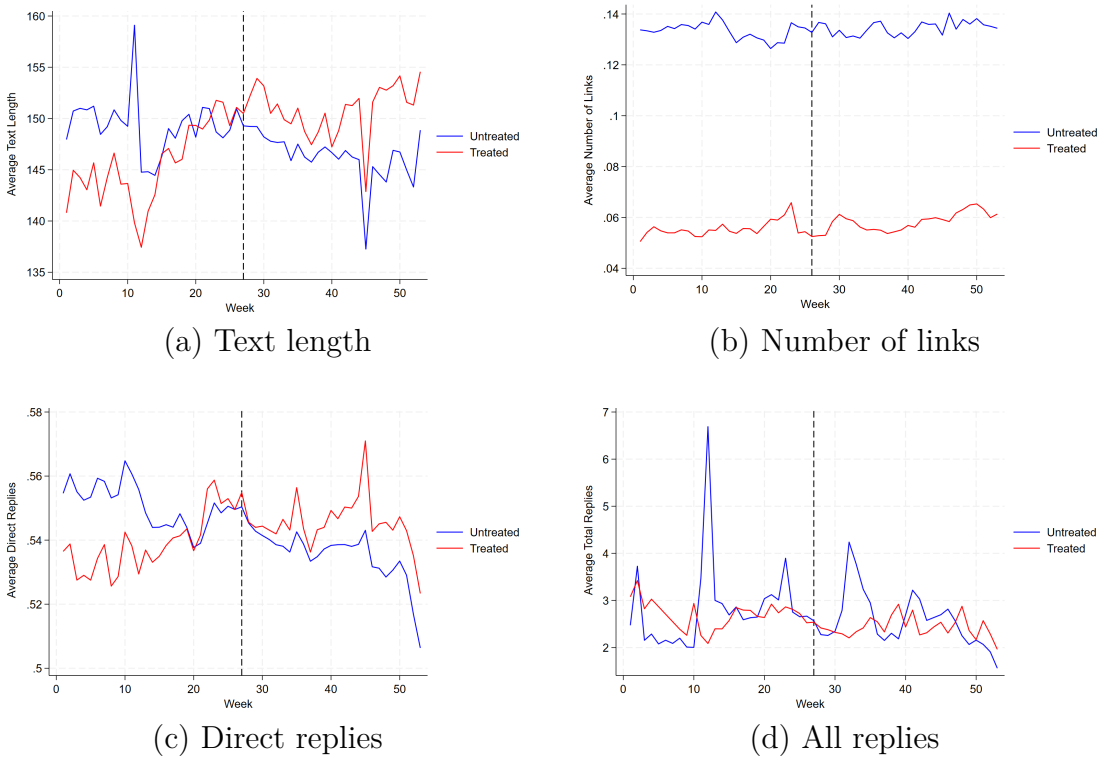
Notes: Figure C.4 illustrates the relationship between degree of policy exposure ($ShareInBan_i$) and hate speech use *outside banned forums* separately for left- and right-wing users. Each observation is a user; incidence of hate speech use is taken as number of slurs per 1000 comments, averaged over weeks of the year. For ease of display incidence of hate speech is log transformed (users who never use hate speech are excluded from the sample). The figure shows that, for accounts which use slurs in the first place, correlation between policy exposure and slur incidence is positive, but modest, and is higher for right-wing users: that is, within this context, being an active user of banned forum does not imply being toxic on the rest of the platform.

Figure C.5: Average number of comments containing slurs by whether comment was a reply to a treated user



Notes: Figure C.5 illustrates the average weekly number of comments containing slurs, separately for treated (left panel) and control (right panel) users. Both panels disaggregate the statistics by whether comment was a reply to treated user. The week before the policy change (*week 26*) is indicated by the dashed vertical line.

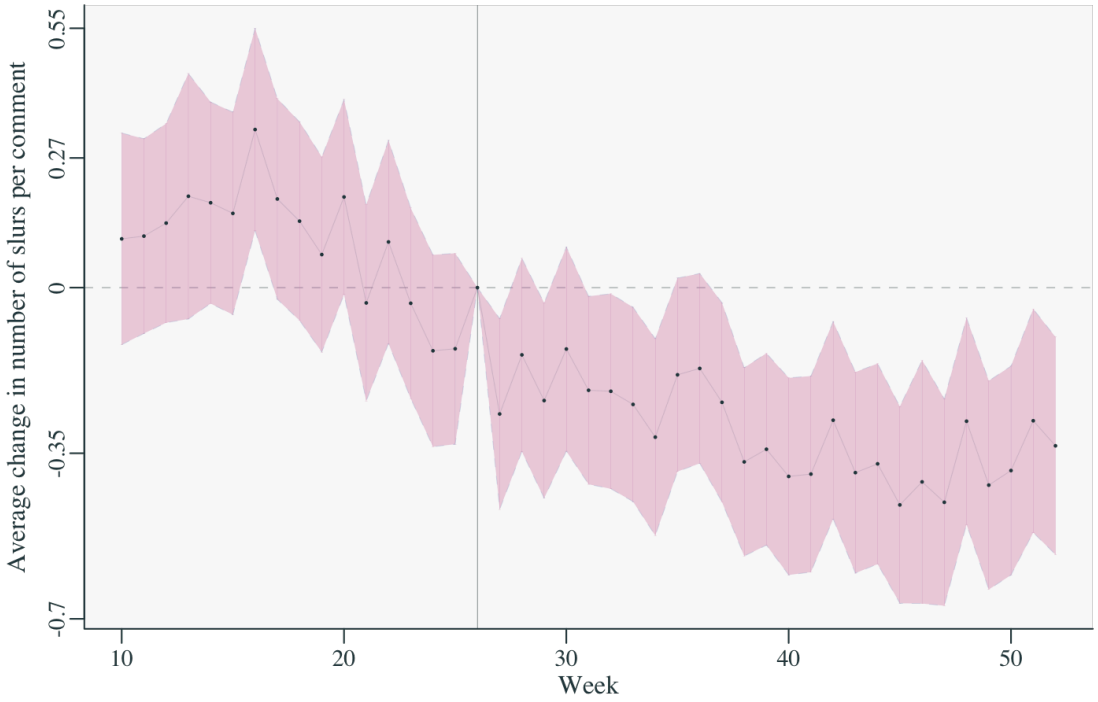
Figure C.6: Measures of engagement: raw data



Notes: Figure C.6 shows the effect of the policy on proxies of engagement quality in the raw data. Panel A displays the average length of a comment, Panel B displays average number of links per comment. Panels C and D display average number of direct and total replies to a comment. The corresponding regression results are presented in Table 3.

Robustness to alternative estimations

Figure C.7: Effect on incidence of hate speech per comment: difference-in-differences estimates



Notes: Figure C.7 shows the results of estimating Equation 5 as an event study with weekly estimates. The outcome variable is the incidence of hate speech per comment. The sample is restricted to users who commented at least 50 times between January and March 2020.The reference week is the week before the policy (week 26) indicated by a vertical line. Confidence intervals are at 95% coverage, errors are clustered at the user level. N=181,686,956.

Table C.1: Effects on incidence of hate speech: difference-in-differences estimates (50 comment cutoff)

	Number of unique slurs per 1000 comments			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i$	-0.341***	-0.278*	-0.294***	-0.214
	(0.060)	(0.109)	(0.077)	(0.162)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	1,632,607			
User-subreddit pairs		85,775,318		
Posts			119,810,427	
Threads				574,042,847
N	1,043,290,141	1,043,290,141	1,043,290,141	1,043,290,141
Dep. var. mean	1.93	1.93	1.93	1.93
R^2	0.04	0.13	0.15	0.58

Notes: Table C.1 presents the results of estimating Equation 5 across all quartiles by policy exposure for users who left at least 50 comments between January and March 2020. For ease of display the dependent variable is multiplied by 1000. We only count unique words due to probable autocorrelation in use of specific language within a comment. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.2: Effects on incidence of hate speech by degree of user exposure to policy change: difference-in-differences estimates

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Q1_{ShareInBan}$	-0.247***	-0.150
	(0.090)	(0.166)
$Post_t \times Q2_{ShareInBan}$	-0.621***	-0.580***
	(0.065)	(0.080)
$Post_t \times Q3_{ShareInBan}$	-0.634***	-0.591***
	(0.085)	(0.093)
$Post_t \times Q4_{ShareInBan}$	-0.314***	-0.300*
	(0.100)	(0.122)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	28,979,600	
User-subreddit pairs		242,242,341
N	1,884,922,792	1,884,922,792
Dep. var. mean	1.82	1.82
R^2	0.05	0.18

Notes: Table C.2 presents the results of estimating Equation 5. Independent variable is interacted with indicator for quartile of policy exposure ($ShareInBan_i$) of the user. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * p < 0.1, ** p < 0.05, *** p < 0.01.

Table C.3: Effects on incidence of hate speech by degree of user exposure to policy change: difference-in-differences estimates (50 comment cutoff)

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Q1_{ShareInBan}$	-0.209*	-0.124
	(0.090)	(0.167)
$Post_t \times Q2_{ShareInBan}$	-0.601***	-0.571***
	(0.070)	(0.084)
$Post_t \times Q3_{ShareInBan}$	-0.600***	-0.598***
	(0.096)	(0.103)
$Post_t \times Q4_{ShareInBan}$	-0.191	-0.240*
	(0.107)	(0.121)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	1,632,607	
User-subreddit pairs		85,775,318
N	1,043,290,141	1,043,290,141
Dep. var. mean	1.93	1.93
R^2	0.04	0.13

Notes: Table C.3 presents the results of estimating Equation 5. Independent variable is interacted with indicator for quartile of policy exposure ($ShareInBan_i$) of the user. Sample is restricted to users who left at least 50 comments between January and March 2020. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.4: Effects on number of comments containing slurs: difference-in-differences estimates

	Number of comments containing slurs (per 1000 comments)			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i$	-0.417***	-0.379***	-0.437***	-0.402***
	(0.032)	(0.028)	(0.037)	(0.043)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	28,979,600			
User-subreddit pairs		242,242,341		
Posts			158,682,911	
Threads				908,670,654
N	1,884,546,944	1,884,546,944	1,884,546,944	1,884,546,944
Dep. var. mean	1.75	1.75	1.75	1.75
R^2	0.05	0.17	0.12	0.51

Notes: Table C.4 presents the results of estimating Equation 5 with dummy for whether a comment contains a slur as dependent variable. For ease of display the resulting coefficients are multiplied by 1000. The effects with number of unique slurs per comment as outcome variable are presented in Table 1. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.5: Effects on the incidence of hate speech: Poisson quasi-maximum likelihood estimates

	Number of unique slurs per 1000 comments		
	(1)	(2)	(3)
$Post_t \times Treated_i$	-0.083***	-0.060	-0.081***
	(0.018)	(0.032)	(0.017)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Post FE			✓
Weeks	53	53	53
Users	1,218,599		
User-subreddit pairs		1,984,508	
Posts			1,953,725
N	765,457,446	262,568,633	320,992,435
Dep. var. mean	1.82	1.82	1.82

Notes: Table C.5 presents the results of estimating Equation 5 with number of unique slurs (incidences of hate speech) per comment as dependent variable using Poisson quasi-maximum likelihood estimator. We count only unique words to avoid overestimating comment toxicity, due to probable autocorrelation in use of specific slurs within a comment. The results are different in magnitude to the normalized coefficients in Table 1 since Poisson QML estimator implicitly drops groups which contain only zeroes. This particularity does not affect the coefficients in fixed effects estimation, but massively changes (here, increases) the estimate of the *average* incidence of hate speech. Therefore, the coefficients are smaller, though still economically and statistically significant. Standard errors in parentheses are clustered at the user level, * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01.

Table C.6: Effects on incidence of hate speech by degree of user exposure to policy change: robustness to trimming

	Number of unique slurs per 1000 comments			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i$	-0.179***	-0.125	-0.160***	-0.120
	(0.045)	(0.079)	(0.050)	(0.093)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	28,979,600			
User-subreddit pairs		242,242,341		
Posts			158,682,911	
Threads				908,670,654
N	1,884,922,792	1,884,922,792	1,884,922,792	1,884,922,792
Dep. var. mean	0.69	0.69	0.69	0.69
R^2	0.02	0.16	0.09	0.50

Notes: Table C.6 presents the results of estimating Equation 5. Independent variable is interacted with indicator for quartile of policy exposure ($ShareInBan_i$) of the user. Dependent variable is is the incidence of slurs trimmed to exclude top-3 most common slurs in each forum. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Heterogeneity by user group

Table C.7: Effects on incidence of hate speech by political leaning of the users: difference-in-differences estimates (50 comment cutoff)

	Number of unique slurs per 1000 comments			
	(1)	(2)	(3)	(4)
$Post_t \times Treated_i \times Left_i$	-0.048	-0.017	0.138***	0.093
	(0.042)	(0.048)	(0.050)	(0.071)
$Post_t \times Treated_i \times Right_i$	-0.351***	-0.288	-0.317***	-0.165
	(0.079)	(0.150)	(0.103)	(0.221)
$Treated_i \times Left_i$			-0.253***	-0.236***
			(0.052)	(0.064)
$Treated_i \times Right_i$			0.848***	0.650***
			(0.050)	(0.091)
Week FE	✓	✓	✓	✓
User FE	✓			
User-subreddit FE		✓		
Post FE			✓	
Thread FE				✓
Weeks	53	53	53	53
Users	1,624,576			
User-subreddit pairs		84,620,442		
Posts			118,749,752	
Threads				570,451,692
N	1,034,500,566	1,034,500,566	1,034,500,566	1,034,500,566
Dep. var. mean	1.93	1.93	1.93	1.93
R^2	0.04	0.13	0.15	0.58

Notes: Table C.7 presents the results of estimating Equation 5. Independent variable is interacted with dummy for political leaning across all quartiles by policy exposure. For ease of display, the dependent variable is multiplied by 1000. Sample is restricted to users who left at least 50 comments in January-March 2020. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.8: Effects on incidence of hate speech by political leaning of the users: robustness to trimming

	Number of unique slurs per 1000 comments		
	(1)	(2)	(3)
$Post_t \times Treated_i \times Left_i$	-0.049*	-0.049	0.041
	(0.021)	(0.026)	(0.025)
$Post_t \times Treated_i \times Right_i$	-0.191***	-0.134	-0.194***
	(0.062)	(0.109)	(0.066)
$Treated_i \times Left_i$			-0.030
			(0.056)
$Treated_i \times Right_i$			0.465***
			(0.068)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Post FE			✓
Weeks	53	53	53
Users	28,963,687		
User-subreddit pairs		240,919,975	
Posts			158,244,079
N	1,875,655,420	1,875,655,420	1,875,655,420
Dep. var. mean	0.69	0.69	0.69
R^2	0.02	0.16	0.09

Notes: Table C.8 presents the results of estimating Equation 5. Independent variable is interacted with dummy for political leaning across all quartiles by policy exposure. Dependent variable is is the incidence of slurs trimmed to exclude top-3 most common slurs in each forum. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.9: Effects on incidence of hate speech by degree of user exposure to policy change and polarity: difference-in-differences estimates (50 comment cutoff)

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Q1_{ShareInBan} \times Left_i$	-0.104	0.008
	(0.073)	(0.089)
$Post_t \times Q1_{ShareInBan} \times Right_i$	-0.201	-0.104
	(0.114)	(0.217)
$Post_t \times Q2_{ShareInBan} \times Left_i$	-0.094	-0.184*
	(0.082)	(0.086)
$Post_t \times Q2_{ShareInBan} \times Right_i$	-0.659***	-0.657***
	(0.082)	(0.108)
$Post_t \times Q3_{ShareInBan} \times Left_i$	-0.039	0.008
	(0.080)	(0.084)
$Post_t \times Q3_{ShareInBan} \times Right_i$	-0.659***	-0.724***
	(0.112)	(0.136)
$Post_t \times Q4_{ShareInBan} \times Left_i$	0.230*	0.229*
	(0.096)	(0.111)
$Post_t \times Q4_{ShareInBan} \times Right_i$	-0.633***	-0.768***
	(0.205)	(0.256)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	1,624,576	
User-subreddit pairs		84,620,442
N	1,034,500,566	1,034,500,566
Dep. var. mean	1.93	1.93
R^2	0.04	0.13

Notes: Table C.9 presents the results of estimating Equation 5. Independent variable is interacted with indicator for quartile of policy exposure ($ShareInBan_i$) of the user and polarity. Sample is restricted to users who left at least 50 comments between January and March 2020. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.10: Effects on incidence of hate speech by political leaning of the users and hate speech category: difference-in-differences estimates

	Number of unique slurs per 1000 comments					
	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
$Post_t \times Treated_i \times Left_i$	-0.005	-0.032	-0.023	0.014	-0.001	0.000
	(0.003)	(0.021)	(0.015)	(0.019)	(0.025)	(0.006)
$Post_t \times Treated_i \times Right_i$	-0.008***	-0.046***	-0.138***	-0.035*	-0.051	-0.041***
	(0.002)	(0.016)	(0.014)	(0.014)	(0.141)	(0.004)
Week FE	✓	✓	✓	✓	✓	✓
User-subreddit FE	✓	✓	✓	✓	✓	✓
Weeks	53	53	53	53	53	53
User-subreddit pairs	240,919,975	240,919,975	240,919,975	240,919,975	240,919,975	240,919,975
N	1,875,279,908	1,875,279,908	1,875,279,908	1,875,279,908	1,875,279,908	1,875,279,908
Dep. var. mean	0.01	0.40	0.20	0.77	0.40	0.04
R^2	0.12	0.12	0.15	0.17	0.23	0.16

Notes: Table C.10 presents the results of estimating Equation 5 . Independent variable is interacted with dummy for political leaning separately by hate speech type. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.11: Effects on incidence of hate speech on core left-wing users: difference-in-differences estimates (not adjusted for selection)

	Number of unique slurs per 1000 comments					
	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
$Post_t \times CoreLeft_i$	-0.001	0.066	0.036	0.011	0.018	0.000
	(0.006)	(0.051)	(0.027)	(0.038)	(0.048)	(0.015)
Week FE	✓	✓	✓	✓	✓	✓
User FE	✓	✓	✓	✓	✓	✓
Weeks	53	53	53	53	53	53
Users	28,810,591	28,810,591	28,810,591	28,810,591	28,810,591	28,810,591
N	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930
Dep. var. mean	0.01	0.40	0.20	0.77	0.40	0.04
R^2	0.02	0.02	0.03	0.04	0.12	0.03

Notes: Table C.11 presents the results of estimating Equation 5 for core left-wing treated users separately by category of slur. For ease of display the dependent variable is multiplied by 1000. Specification includes only user fixed-effects, estimating equation equivalent to Column 1 of Table 1. We only count unique words due to probable autocorrelation in use of specific language within a comment. Equivalent estimates on the subsample of users who left at least 50 comments in January-March 2020 are presented in Table C.13. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.12: Effects on incidence of hate speech on core left-wing users: difference-in-differences estimates

	Number of unique slurs per 1000 comments					
	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
$Post_t \times CoreLeft_i$	-0.003	0.086	0.046	-0.008	0.035	-0.002
	(0.007)	(0.060)	(0.034)	(0.050)	(0.056)	(0.013)
Week FE	✓	✓	✓	✓	✓	✓
User-subreddit FE	✓	✓	✓	✓	✓	✓
Weeks	53	53	53	53	53	53
User-subreddit pairs	231,902,811	231,902,811	231,902,811	231,902,811	231,902,811	231,902,811
N	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930	1,790,619,930
Dep. var. mean	0.01	0.40	0.20	0.77	0.40	0.04
R^2	0.12	0.12	0.15	0.17	0.24	0.17

Notes: Table C.12 presents the results of estimating Equation 5 for core left-wing treated users separately by category of slur. For ease of display the dependent variable is multiplied by 1000. We only count unique words due to probable autocorrelation in use of specific language within a comment. Equivalent estimates on the subsample of users who left at least 50 comments in January-March 2020 are presented in Table C.14. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.13: Effects on incidence of hate speech on core left-wing users: difference-in-differences estimates (50 comment cutoff, not adjusted for selection)

	Number of unique slurs per 1000 comments					
	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
$Post_t \times CoreLeft_i$	-0.001	0.113*	0.061*	0.015	0.044	-0.001
	(0.006)	(0.055)	(0.027)	(0.040)	(0.052)	(0.015)
Week FE	✓	✓	✓	✓	✓	✓
User FE	✓	✓	✓	✓	✓	✓
Weeks	53	53	53	53	53	53
Users	1,536,709	1,536,709	1,536,709	1,536,709	1,536,709	1,536,709
N	954,427,016	954,427,016	954,427,016	954,427,016	954,427,016	954,427,016
Dep. var. mean	0.01	0.40	0.20	0.77	0.40	0.04
R^2	0.00	0.01	0.01	0.01	0.13	0.01

Notes: Table C.13 presents the results of estimating Equation 5 for core left-wing treated users separately by category of slur. For ease of display the dependent variable is multiplied by 1000. Specification includes only user fixed-effects, estimating equation equivalent to Column 1 of Table 1. We only count unique words due to probable autocorrelation in use of specific language within a comment. The sample is restricted to users who left at least 50 comments in January-March 2020. Equivalent estimates on the whole sample of core left-wing users are presented in Table C.11. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.14: Effects on incidence of hate speech on core left-wing users: difference-in-differences estimates (50 comment cutoff)

	Number of unique slurs per 1000 comments					
	Antisemitic	Disability	Homophobic	Misogynistic	Racial/ethnic	Transphobic
$Post_t \times CoreLeft_i$	-0.002	0.119	0.061	-0.001	0.054	-0.002
	(0.008)	(0.063)	(0.033)	(0.053)	(0.059)	(0.012)
Week FE	✓	✓	✓	✓	✓	✓
User-subreddit FE	✓	✓	✓	✓	✓	✓
Weeks	53	53	53	53	53	53
User-subreddit pairs	76,824,252	76,824,252	76,824,252	76,824,252	76,824,252	76,824,252
N	954,427,016	954,427,016	954,427,016	954,427,016	954,427,016	954,427,016
Dep. var. mean	0.01	0.40	0.20	0.77	0.40	0.04
R^2	0.07	0.09	0.10	0.11	0.21	0.10

Notes: Table C.14 presents the results of estimating Equation 5 for core left-wing treated users separately by category of slur. For ease of display the dependent variable is multiplied by 1000. We only count unique words due to probable autocorrelation in use of specific language within a comment. The sample is restricted to users who left at least 50 comments in January-March 2020. Equivalent estimates on the whole sample of core left-wing users are presented in Table C.12. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.15: Effects on incidence of hate speech by banned forum type user visited most: difference-in-differences estimates

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Treated_i \times AltRight_i$	-0.568	-0.278
	(0.304)	(0.637)
$Post_t \times Treated_i \times AntiChina_i$	-0.308	-0.456*
	(0.213)	(0.229)
$Post_t \times Treated_i \times TheDonald_i$	-0.295***	-0.254***
	(0.050)	(0.052)
$Post_t \times Treated_i \times Left_i$	-0.084*	-0.039
	(0.040)	(0.047)
$Post_t \times Treated_i \times Offensive_i$	-0.246***	-0.207***
	(0.044)	(0.040)
$Post_t \times Treated_i \times Other_i$	-0.826***	-0.919***
	(0.194)	(0.120)
$Post_t \times Treated_i \times Sexual_i$	-1.105***	-0.843***
	(0.233)	(0.199)
$Post_t \times Treated_i \times Transphobic_i$	-0.281***	-0.258*
	(0.086)	(0.102)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	28,979,600	
User-subreddit pairs		242,242,341
N	1,884,922,792	1,884,922,792
Dep. var. mean	1.82	1.82
R^2	0.05	0.18

Notes: Table C.15 presents the results of estimating Equation 5. Independent variable is interacted with indicator for type of banned forum a treated user was active in prior to the policy change across all quartiles by policy exposure. For ease of display, the dependent variable is multiplied by 1000. The effects by left and right leaning are presented in Table 2. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Heterogeneity by local language norm

Table C.16: Effects on incidence of hate speech by quartile of slur use within subreddit: difference-in-differences estimates

	Number of unique slurs per 1000 comments		
	(1)	(2)	(3)
$Post_t \times Treated_i \times Q_0$	-1.203***	0.625***	0.083
	(0.066)	(0.073)	(0.076)
$Post_t \times Treated_i \times Q_1$	-1.873***	0.082***	-0.092***
	(0.035)	(0.020)	(0.016)
$Post_t \times Treated_i \times Q_2$	-0.981***	-0.073***	-0.286***
	(0.035)	(0.027)	(0.027)
$Post_t \times Treated_i \times Q_3$	2.001***	-1.097***	-1.214***
	(0.083)	(0.082)	(0.121)
$Post_t \times Treated_i \times Q_4$	13.714***	-4.984	-1.157
	(2.600)	(5.542)	(4.470)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Post FE			✓
Weeks	53	53	53
Users	28,516,722		
User-subreddit pairs		238,231,985	
Posts			155,535,656
N	1,867,164,518	1,867,164,518	1,867,164,518
Dep. var. mean	1.82	1.82	1.82
R^2	0.05	0.17	0.13

Notes: Table C.16 presents the results of estimating Equation 5. Independent variable is interacted with indicator for the quartile of slur incidence of subreddit where the comment has been left. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.17: Effects on incidence of hate speech by quartile of slur use within subreddit: left-wing users

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Treated_i \times Q_0$	-0.318***	0.561***
	(0.112)	(0.177)
$Post_t \times Treated_i \times Q_1$	-0.938***	0.153***
	(0.041)	(0.033)
$Post_t \times Treated_i \times Q_2$	-0.387***	0.012
	(0.046)	(0.057)
$Post_t \times Treated_i \times Q_3$	1.608***	-0.267
	(0.103)	(0.159)
$Post_t \times Treated_i \times Q_4$	6.894***	-3.437***
	(0.677)	(1.091)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	28,371,066	
User-subreddit pairs		229,475,255
N	1,788,429,239	1,788,429,239
Dep. var. mean	1.82	1.82
R^2	0.05	0.18

Notes: Table C.17 presents the results of estimating Equation 5 on the subsample of left-wing treated users. Independent variable is interacted with indicator for the quartile of slur incidence of subreddit where the comment has been left. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * p < 0.1, ** p < 0.05, *** p < 0.01.

Table C.18: Effects on incidence of hate speech by quartile of slur use within subreddit: right-wing users

	Number of unique slurs per 1000 comments	
	(1)	(2)
$Post_t \times Treated_i \times Q_0$	-1.153***	0.546***
	(0.067)	(0.077)
$Post_t \times Treated_i \times Q_1$	-1.809***	0.058***
	(0.036)	(0.021)
$Post_t \times Treated_i \times Q_2$	-0.933***	-0.064*
	(0.039)	(0.031)
$Post_t \times Treated_i \times Q_3$	1.823***	-1.201***
	(0.093)	(0.093)
$Post_t \times Treated_i \times Q_4$	12.474***	-3.227
	(3.787)	(8.125)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	28,465,454	
User-subreddit pairs		235,210,157
N	1,841,370,843	1,841,370,843
Dep. var. mean	1.82	1.82
R^2	0.05	0.17

Notes: Table C.18 presents the results of estimating Equation 5 on the subsample of right-wing treated users. Independent variable is interacted with indicator for the quartile of slur incidence of subreddit where the comment has been left. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * p < 0.1, ** p < 0.05, *** p < 0.01.

Table C.19: Effects on incidence of hate speech in replies to hateful comments: difference-in-differences estimates

	Number of unique slurs per 1000 comments	
	(1)	(2)
$ParentSlur_c$	96.545***	87.891***
	(4.498)	(4.647)
$Post_t \times ParentSlur_c$	-0.308	-0.950
	(1.062)	(0.988)
$ParentSlur_c \times Treated_i$	17.827***	9.583*
	(6.503)	(4.869)
$Post_t \times Treated_i$	-0.319***	-0.352***
	(0.036)	(0.034)
$Post_t \times ParentSlur_c \times Treated_i$	-11.514***	-4.788
	(4.164)	(2.587)
Week FE	✓	✓
User FE	✓	
User-subreddit FE		✓
Weeks	53	53
Users	19,554,468	
User-subreddit pairs		132,365,277
N	950,894,365	950,894,365
Dep. var. mean	1.82	1.82
R^2	0.07	0.19

Notes: Table C.19 presents the results of estimating Equation 5. Independent variable is interacted with dummy for whether a comment is itself a reply to a comment containing slur. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.20: Effects on incidence of hate speech in replies to slurs and treated users: difference-in-differences estimates

	Number of unique slurs per 1000 comments					
	(1)	(2)	(3)	(4)	(5)	(6)
$ParentSlur_c$	96.545***	87.891***			96.180***	87.916***
	(4.498)	(4.647)			(4.517)	(4.691)
$Post_t \times ParentSlur_c$	-0.308	-0.950			-0.145	-0.775
	(1.062)	(0.988)			(1.062)	(0.999)
$ParentSlur_c \times Treated_i$	17.827***	9.583*			12.310***	9.133
	(6.503)	(4.869)			(4.700)	(4.868)
$Post_t \times Treated_i$	-0.319***	-0.352***	-0.262***	-0.369***	-0.213***	-0.313***
	(0.036)	(0.034)	(0.040)	(0.034)	(0.036)	(0.033)
$Post_t \times ParentSlur_c \times Treated_i$	-11.514***	-4.788			-7.234***	-5.393*
	(4.164)	(2.587)			(2.447)	(2.226)
$ParentTreated_c$			0.770***	0.299***	0.606***	0.214***
			(0.032)	(0.019)	(0.032)	(0.019)
$Post_t \times ParentTreated_c$			-0.331***	-0.211***	-0.246***	-0.152***
			(0.021)	(0.022)	(0.022)	(0.022)
$ParentTreated_c \times Treated_i$			1.147***	0.186***	0.960***	0.118*
			(0.067)	(0.060)	(0.066)	(0.059)
$Post_t \times ParentTreated_c \times Treated_i$			-0.785***	-0.324***	-0.616***	-0.222*
			(0.094)	(0.089)	(0.092)	(0.088)
Week FE	✓	✓	✓	✓	✓	✓
User FE	✓		✓		✓	
User-subreddit FE		✓		✓		✓
Weeks	53	53	53	53	53	53
Users	19,554,468		19,357,324		19,357,324	
User-subreddit pairs		132,365,277		130,146,466		130,146,466
N	950,894,365	950,894,365	915,561,017	915,561,017	915,561,017	915,561,017
Dep. var. mean	1.82	1.82	1.82	1.82	1.82	1.82
R^2	0.07	0.19	0.04	0.17	0.05	0.18

Notes: Table C.20 presents the results of estimating Equation 5. Independent variable is interacted with dummies for whether a comment is itself a reply to a comment containing slur or a reply to another treated user. For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.21: Effects on incidence of hate speech in replies to treated users: difference-in-differences estimates

	Number of unique slurs per 1000 comments		
	(1)	(2)	(3)
$ParentTreated_c$	0.767***	0.296***	-0.642
	(0.031)	(0.019)	(912.886)
$Post_t \times ParentTreated_c$	-0.328***	-0.209***	0.010
	(0.021)	(0.022)	(0.080)
$ParentTreated_c \times Treated_i$	1.121***	0.165***	1351.081
	(0.067)	(0.060)	(4042.523)
$Post_t \times Treated_i$	-0.274***	-0.377***	-0.371***
	(0.039)	(0.034)	(0.092)
$Post_t \times ParentTreated_c \times Treated_i$	-0.763***	-0.304***	-0.146
	(0.094)	(0.089)	(0.606)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Pair FE			✓
Weeks	53	53	53
Users	19,401,553		
User-subreddit pairs		130,598,787	
Pairs			529,155,670
N	919,121,114	919,121,114	915,561,017
Dep. var. mean	1.82	1.82	1.82
R^2	0.04	0.17	0.56

Notes: Table C.21 presents the results of estimating Equation 5. Independent variable is interacted with indicator for whether a comment is a reply to a treated user For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Table C.22: Effects on incidence of hate speech within pairs of treated users: difference-in-differences estimates

	Slurs per 1000 comments	**Comments containing slurs**
$Treated_i$	1.007***	0.969***
	(0.257)	(0.244)
$Post_t \times Treated_i$	-0.442***	-0.423***
	(0.070)	(0.065)
Week FE	✓	✓
Pair FE	✓	✓
Weeks	53	53
Pairs	530,379,877	530,379,877
N	942,745,961	942,745,961
Dep. var. mean	1.82	1.75
R^2	0.54	0.54

Notes: Table C.22 presents the results of estimating Equation 5 . For ease of display, the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

Comment toxicity

Table C.23: Effects on use of strong language: difference-in-differences estimates

	Use of strong language		
	(1)	(2)	(3)
$Post_t \times Treated_i$	-2.403***	-1.602***	-1.021
	(0.392)	(0.394)	(0.964)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Post FE			✓
Weeks	53	53	53
Users	28,979,600		
User-subreddit pairs		242,242,341	
Posts			158,682,911
N	1,884,922,792	1,884,922,792	1,884,922,792
Dep. var. mean	108.13	108.13	108.13
R^2	0.14	0.26	0.17

Notes: Table C.23 presents the results of estimating Equation 5 with severity adjusted strong language use score as dependent variable. For ease of display the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * p $<$ 0.1, ** p $<$ 0.05, *** p $<$ 0.01.

Table C.24: Effects on use of abusive language: difference-in-differences estimates

	Use of abusive language		
	(1)	(2)	(3)
$Post_t \times Treated_i$	0.328***	0.241***	-0.349***
	(0.072)	(0.073)	(0.131)
Week FE	✓	✓	✓
User FE	✓		
User-subreddit FE		✓	
Post FE			✓
Weeks	53	53	53
Users	28,979,600		
User-subreddit pairs		242,242,341	
Posts			158,682,911
N	1,884,922,792	1,884,922,792	1,884,922,792
Dep. var. mean	8.90	8.90	8.90
R^2	0.07	0.19	0.12

Notes: Table C.24 presents the results of estimating Equation 5 with severity adjusted abusive language use score as dependent variable. For ease of display the dependent variable is multiplied by 1000. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

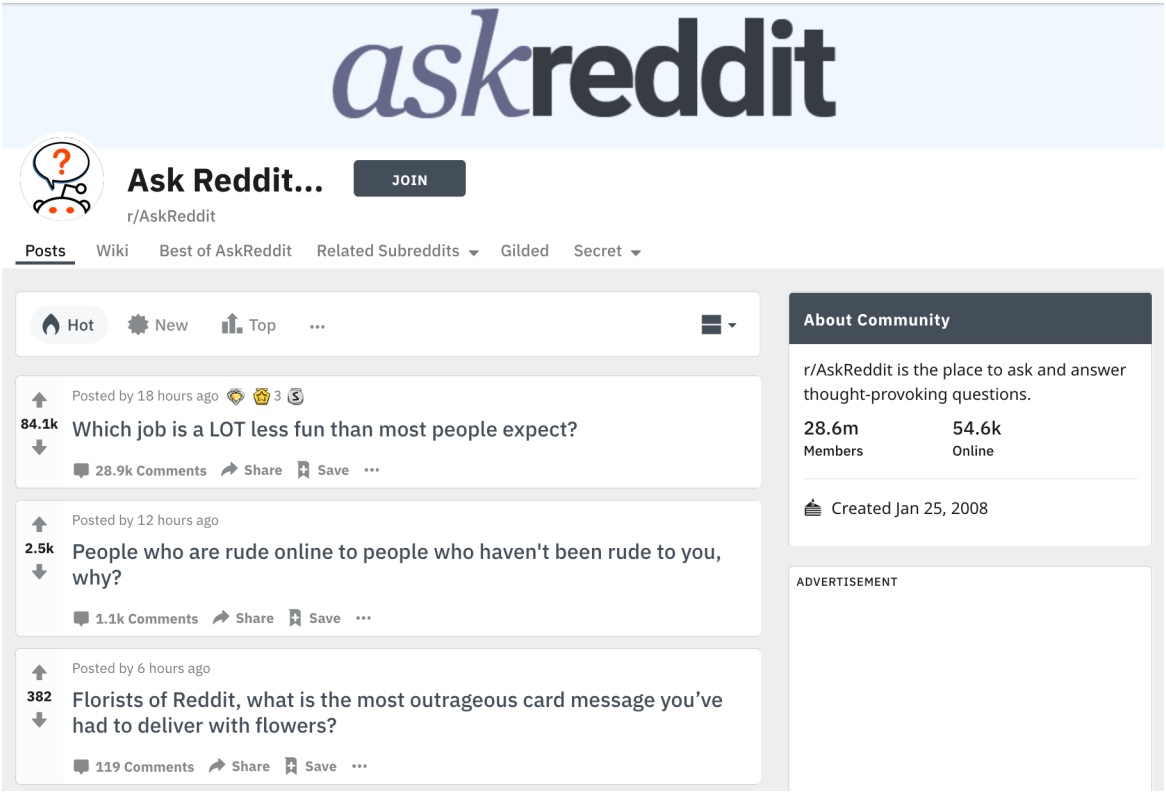
Table C.25: Effects on use of toxic language by political leaning of the users: difference-in-differences estimates (severity adjusted)

	Comment toxicity score	
	Strong language score	Abusive language score
$Post_t \times Treated_i \times Left_i$	-3.182***	-0.030
	(0.779)	(0.146)
$Post_t \times Treated_i \times Right_i$	-0.428	0.433***
	(0.393)	(0.081)
Week FE	✓	✓
User-subreddit FE	✓	✓
Weeks	53	53
User-subreddit pairs	240,919,975	240,919,975
N	1,875,655,420	1,875,655,420
Dep. var. mean	108.13	8.90
R^2	0.26	0.18

Notes: Table C.25 presents the results of estimating Equation 5 with severity adjusted incidence of toxic language as dependent variable. Independent variable is interacted with dummy for political leaning across all quartiles by policy exposure. For ease of display, the dependent variable is multiplied by 1000. The effects for use of hate speech are presented in Table 2. Standard errors in parentheses are clustered at the user level, * $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$.

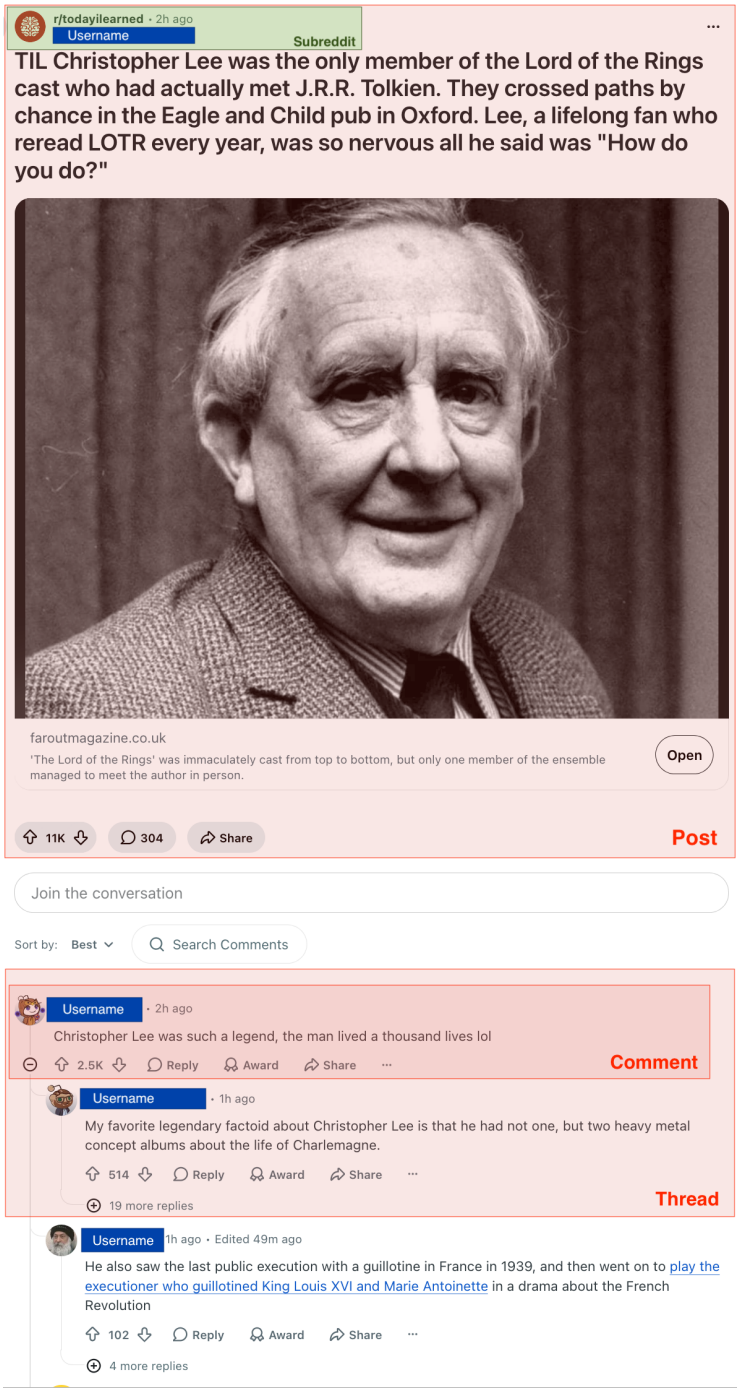
D Platform overview

Figure D.1: User interface of Reddit: subreddit



Notes: Figure D.1 shows the interface of an individual subreddit as it looked in June 2020: here, `r/AskReddit`, by far the largest forum on the platform at the time of the policy change. The interface of the website has changed over the years, but the the main components have remained the same (user names in this example are not visible due to the way Wayback Machine saves snapshots, but they are normally displayed along with the post). Users can leave posts within *subreddits*, upvote or downvote existing posts or click on posts to engage in a specific conversation further. Sum of upvotes and downvotes affects the position of the post relative to other posts within a forum and determines whether a post reaches *front page of Reddit*. Reddit automatically sorts posts based on the sum of upvotes and downvotes, but users are free to sort by chronological order as well. Source: https://web.archive.org/web/20200629193736/https://www.reddit.com/r/AskReddit/ (last accessed 9/9/2026 via Wayback Machine).

Figure D.2: User interface of Reddit: labelled elements of a post



Notes: Figure D.2 shows an example of a public post on Reddit from *subreddit* `r/todayilearned`, one of the largest communities on Reddit. The interface of the website has changed over the years, but the the main components have remained the same. Users can leave posts within subreddits, comment on the posts themselves or comments of other users. *Threads* are subtrees within comments section reflecting separate conversation chains. Additionally, in most cases users can upvote or downvote a post or a comment, affecting the position of comment relative to other comments on the same level (comments which are replies to the same parent comment or post) and position of post within a forum. Reddit automatically sorts comments based on the sum of upvotes and downvotes, but users are free to sort by chronological order as well. Source: https://www.reddit.com/r/todayilearned/comments/1wbm7e5/til_christopher_lee_was_the_only_member_of_the/ (last accessed 9/9/2026).