Hate in the Time of Algorithms: Evidence from a Large-Scale Experiment on Online Behavior

Aarushi Kalra*

Job Market Paper

January 25, 2025

(Please click here for the most recent version)

Abstract

As social media usage reaches record highs, personalization algorithms risk radicalizing users by reinforcing existing beliefs. However, evidence on how algorithms and user behavior jointly shape harmful online engagement is limited. In this paper, I conduct an individually randomized experiment with 8 million users of a prominent TikTok-like platform in India, replacing the feed-ranking algorithm with random content delivery. I focus on hateful content targeting minority groups, given its prominence on Indian social media and establish a trade-off: random post recommendation lowers exposure to anti-minority ("toxic") content by 27%, but at a substantial cost to the platform as overall platform usage falls by 35%. Strikingly, treated users share a larger proportion of the toxic posts they view, mitigating the decline in the number of toxic posts shared from the platform. Users with a higher interest in toxic content at baseline drive this result as they seek out posts the algorithm does not show them. I rationalize these findings with a model of a revenue-driven algorithm that faces heterogeneous users choosing which posts to consume. Counterfactual simulations evaluate alternative interventions that target toxicity in the algorithm's recommendations. Finally, I collect survey evidence to trace users' behavior beyond the platform and show that the most affected users substitute away to other platforms. These results underscore the limits of piecemeal algorithmic regulation intended to moderate harmful content online.

Keywords: AI, Digital Platforms, Algorithms, Toxicity, Development

*Department of Economics, Brown University, Providence, RI, 02906 (email: aarushi_kalra@brown.edu).
I am grateful to Andrew Foster, Brian Knight, Daniel Björkegren, Peter Hull, and Stelios Michalopoulos, for their continued guidance and support. This project has greatly benefitted from helpful discussions with Pedro Dal Bo, Bryce Steinberg, Jesse Bruhn, Neil Thakral, Ro'ee Levy, Matthew Pecenco, Lorenzo Lagos, Elisa Macchi, Martin Mattsson and seminar participants at Brown University. I thank Ahad Bashir and Farrukh Zaidi for excellent research assistance. The experiment was preregistered on the AEA RCT registry, ID AEARCTR-0010933. Protocols for survey data collection were approved by the Institutional Review Board at Brown University. This project was generously supported by the NSF Dissertation Research Improvement Grant, the Weiss Family Fund for Research in Development Economics, the Orlando Bravo Center for Economic Research and the Saxena Center for Contemporary South Asia.

1 Introduction

Social media platforms engage 64% of the world's population, raising concerns about the potential exposure of 5 billion users to harmful unmoderated content (Acemoglu et al., 2023). Personalization algorithms designed to maximize engagement can create radicalizing echo chambers (Sunstein, 2018) that have been linked with physical violence against minorities (Müller and Schwarz, 2021). These concerns have engendered proposals for a regulatory response to diversify content feeds by disabling algorithmic personalization.[1] While such content-moderation policies can reduce exposure to harmful content, little is known about their effectiveness which can be limited if users' online behavior is unresponsive to viewing diverse information (Hosseinmardi et al., 2024). Further, industry-wide policy adoption is uncertain due to costs imposed on revenue-maximizing platforms and social media users who value the algorithm's recommendations (Acemoglu, 2021; Kasy, forthcoming).

In this paper, I conduct an individually randomized experiment with 8 million social media users to study the causal effect of "switching-off" personalization algorithms on online engagement, focusing on interactions with "toxic" content.[2] I develop a partnership with an Indian social media platform (henceforth, SM), that has a user base equivalent to the combined population of Germany, France, and the UK–about 200 million users–to study the impact of diversifying content feeds by randomizing exposure to different kinds of posts.[3] The recommendation algorithms employed by this TikTok-like platform I collaborate with, akin to the algorithms typically used by YouTube, Netflix, Amazon or Spotify, optimize for the time users spend online based on their previous engagement with similar content. This feed-ranking algorithm remains unchanged for 3% of SM's user base, 6 million users operating the application in over a dozen regional languages, who were randomly assigned to the control group. I directly intervene on the personalization algorithm in a 1% random sample–2 million users–with randomized content delivery to find a reduction in the absolute number of toxic posts shared, and a positive effect on the relative quantities of harmful shares which is shown to undo some of the gains from the intervention.

Studying the causal effects of algorithmic regulation is challenging for several reasons. First, user behavior and personalized recommendations are endogenous as algorithms "learn" from past interactions to maximize engagement. Prior on-platform experimental research

[1] See `https://shorturl.at/WKWPU` for minutes of the Subcommittee on Privacy, Technology, and the Law convened under the US Senate Committee on the Judiciary.

[2] Toxicity, as defined by Google's Perspective API, measures a post's potential harm as it scores comments on a scale of 0 to 1 based on their likelihood to make someone leave a discussion. This API is used by organizations like the New York Times and in academic research (Jiménez Durán et al., 2024).

[3] The DUA between Brown University and SM does not impose restrictions on the publication of the results of the study due to the anonymization of firm identity and relevant measures to protect user privacy.

partially addresses these concerns, finding that algorithmic interventions are ineffective in reducing polarization (Guess et al., 2023a,b; Nyhan et al., 2023). However, these experiments may not isolate demand for polarizing content because the interventions employed, such as recommending content in reverse chronological order, still depend on social networks that reflect user preferences. Second, opt-in requirements in these closely related studies likely introduce selection bias. Third, supply-side incentives to "self-regulate" are opaque as feed-ranking algorithms optimize over unknown objectives that may not align with the regulator's goals of minimizing harm (Jiménez Durán, 2022). Further, platforms are reluctant to even experimentally remove posts based on subjective assessments of harm as they want to avoid being accused of political bias or suppressing free speech (Kominers and Shapiro, 2024). This makes it hard to collect evidence that can inform design of content moderation policies.

My study offers the largest-scale experimental evidence to date on the effects of personalization algorithms. It is conducted outside US and Europe in a context where institutional and cultural constraints to regulate the ICT sector, that constitutes 13% of India's GDP, are distinct and understudied (ITA, 2024; Blair et al., 2024). I address the identification challenge cited above by replacing algorithmically curated feeds with randomly picked content. The intervention has limited general equilibrium effects as the incentives of content creators remain unchanged, and on-platform interactions do not depend on users networks on SM, thus minimizing spillovers.[4] The experimental sample is not selected as users consented through the platform's terms of service. High-frequency engagement data characterize engagement decisions and survey outcomes evaluate behavior beyond SM. I model the optimization problems of the platform and users to test the impact of supply- and demand-side factors on engagement with harmful content, and evaluate counterfactual moderation policies that target certain users or posts but are not feasible to implement in the field.

The first set of results focuses on treatment intensity that varied among users, particularly affecting those who viewed the least or the most toxic content at baseline. Random content delivery enhanced the diversity of user feeds as treated users encountered fewer posts of the types they were used to consuming. That is, some users had, through prior platform-engagement, tuned their algorithm to receive large amounts of toxic posts ex-ante, and so were expected to receive more toxic content at baseline and under the control condition. However, users with the highest exposure to toxic content at baseline experienced the largest reductions in toxic exposure upon being treated, 49% lower than the 38 toxic posts viewed over a period of one month by control users in this group. On the other hand, control users

[4]Supplier incentives are unchanged because the treatment is assigned at the user level only for 1% of the platform's user base. Control users are unaffected by online behavior of treated users as the control algorithm's recommendations on SM minimally depend on social networks. This is because content is typically shared on other platforms like WhatsApp and users typically do not follow each other on SM.

with the lowest baseline exposure to toxic content viewed 10 toxic posts on average, whereas treated users in this group viewed 3 more toxic posts during the same time period. On average, the intervention reduced the number of toxic posts viewed in one month by 27%.

Similarly, the intervention altered the feed composition with respect to other types of posts, such as religious or romantic content. I focus on the intervention's impact on engagement with toxic content as I show that the percentage reduction in exposure to this content category is most salient across various genres and topics, among top users of the platform. This focus on toxicity also addresses a key policy concern, especially in India where online misinformation has been linked to deaths fueled by anti-Muslim hate crimes (Banaji et al., 2019). I identify posts that verbally attack or threaten India's Muslim minority using posts deemed political as per SM's internal classification, and toxic or abusive as per the Google-developed Perspective API. The latter algorithm scores a "phrase based on the perceived impact the text may have in conversation."

My second set of results deals with overall platform usage. Random content delivery decreased the total time an average user spent online over a period of one month by 35%, around 5 minutes per day. The average treated user viewed 35 fewer posts compared to control users who viewed 247 posts of any variety in the same time period, a 14% decrease in activity by this measure. On the extensive margin, the average decrease in the number of logins was only 6% when the average control user logged in 22 times in a month.

The largest decrease in overall usage is due to users who viewed the most toxic content at baseline, with a 23% reduction in the total number of posts viewed. These users also reported spending more time on other platforms in my endline survey consisting of over 8,000 randomly selected users from the experimental sample. This means that while platforms are unlikely to self-regulate in the short run, such content moderation policies may be effective in driving out problematic users from one platform, but potentially harming users on others. This result may also raise concerns about differential attrition as treated users with higher affinity towards toxic content were more likely to leave the platform. However, I show that the Lee bounds on treatment effects are tightly estimated (Lee, 2009).

My third set of results considers user engagement with toxic content. The average treated user reduced the number of toxic posts shared by 20% over one month. This reduction is smaller than the average decrease in the number of toxic posts viewed in the short run. Users with the highest exposure to toxic content drive this decrease in toxic engagement on the platform as they reduced the number of toxic posts shared by 34%, where toxic posts constituted 3% of the control users' total shares in a month. On the other hand, users with low toxic exposure at baseline exhibit inelastic sharing behavior and do not share more toxic posts even as they are served a larger number of toxic posts upon bring treated.

My fourth set of results uncovers the degree of malleability in user behavior. Although random content delivery reduced the number of toxic posts viewed and shared, the sharing response to decreased toxic exposure is blunted by the increased rate of sharing such posts. This is because the intervention led to an 18% increase in the probability of sharing toxic posts conditional on viewing them over one month. The implication is that the reduction in the number of toxic posts shared would have been larger if users were more responsive to diversified feeds. This means that users seek out posts that align with their tastes even when they are not readily served to them. This claim of immalleable user behavior is supported by survey evidence: political attitudes are indistinguishable between treatment and control users in the post-intervention period. The effects are precisely estimated as the survey design enables me to rule out even modest changes in attitudes, as small as a 1 pp change.

To rationalize these results, I propose and estimate a model of user behavior on social media platforms. In this model, time spent on the platform is endogenous, users choose the posts to share in order to balance preferences for content consumption and their social-image concerns, and platforms maximize engagement by choosing the proportion of toxic posts to show a user during her session. I predict observed behavior in an equilibrium where users receive both consumption utility from viewing and public recognition utility from sharing posts that are perceived to be socially acceptable. Users' perception of social norms is informed by the algorithmic feed, and I show that the algorithm's incentives lead to value-misalignment between regulators and platforms in the short run. This microfoundation provides an estimation strategy for key behavioral parameters that cannot be directly estimated with the experimental data, such as the elasticity of sharing with respect to exposure.[5] These estimates are then used to simulate the effects of counterfactual policy proposals.

Bringing the model to the data, I find that a 1% decrease in exposure to toxic content decreases toxic sharing only by 0.16%. This elasticity measure, identified using a steady state condition, highlights the limited role of the influence of content exposure on online behavior, and shows that user behavior is not malleable in the short run. Therefore, behavior is largely driven by pre-existing user tastes that are represented by their content exposure in equilibrium at baseline. Simulated policy counterfactuals suggest that interventions targeting toxic posts, such as diversifying feeds of users who have shown a higher proclivity towards such content, are ineffective in changing behavior if user responses are counter to the policy. However, as I show, a combination of diversified and customized feeds can be used to lower the dissemination of toxic content while minimizing the risk of losing users.

[5]This is because in replacing the personalization algorithm with random content delivery, treated users are exposed to a random draw of posts from an average user's feed over time. By the law of large numbers, the average proportion of user feeds consisting of toxic posts is constant for treated users. This means that the variation in "random" exposure is insufficient to identify this influence parameter.

These results have important policy implications for content moderation on digital platforms especially in scarcely regulated environments. First, platforms will not willingly moderate content by diversifying feeds because they lose user engagement. Second, targeting this intervention to users who are likely to share toxic content can help reduce the spread of harmful content as these users disengage with the platform. Third, substitutability among platforms necessitates cross-platform regulation. Fourth, limited malleability in behavior suggests that blanket regulations targeting algorithms may not be as effective as hoped for.

The rest of the paper is organized as follows. Section 2 presents background of the study as well as the administrative, experimental, and survey data sources. Section 3 outlines the experiment's design and presents descriptive statistics. Section 4 presents the main empirical results. Section 5 introduces the theoretical framework, and the model parameters are estimated in Section 6. Section 7 concludes.

1.1 Relation to the Literature

This paper contributes to three strands of the literature. The first strand examines the role of new communication technologies in aggravating political divisions. In collaboration with Meta, Guess et al. (2023a,b) find that on-platform interventions were not effective in reducing polarization, and users seek out like-minded content sources (Nyhan et al., 2023). While my results on overall platform usage are consistent with closely related studies (Beknazar-Yuzbashev et al., 2022), I complement the literature by showing that users with a higher exposure to toxic content at baseline are more likely to disengage with the platform, thus, driving the decrease in engagement with toxic content. Furthermore, India, despite being the second-largest market for digital platforms, and facing a concerning political climate with weak regulatory institutions, is severely understudied. I study a period of "calm" in this context, as opposed to the election period as in the Meta studies, further enhancing the generalizability of the results (Bagchi et al., 2024). Finally, my model-based counterfactuals simulate the effects of policies that are difficult to implement in the field, such as a direct censorship of posts and users.

This paper also contributes to a rich literature finding strong effects of media bias on political polarization (DellaVigna and Kaplan, 2007). Although social media differs from traditional media in some key aspects (lower entry costs for creators and content personalization), existing research shows that supply factors influence behavior (Chiang and Knight, 2011). I build on these studies to show that demand for slanted information drives outcomes (Gentzkow and Shapiro, 2010; Martin and Yurukoglu, 2017) even when supply is endogenous due to algorithms. In so doing, the paper joins a growing literature on the welfare effects

of social media (Allcott et al., 2020; Bursztyn et al., 2023; Brynjolfsson et al., 2024). This literature also shows that social media usage can positively impact political outcomes, and encourage persistent civic engagement, especially among the underrepresented (Bursztyn et al., 2021; Zhuravskaya et al., 2020). However, direct evidence on the mechanisms driving behavioral responses due to (social) media bias is scarce, especially on TikTok-like platforms (Aridor et al., 2024). This paper shows that although users can be influenced by social media, pre-existing biases dominate behavior.

Finally, this paper adds to a literature on the consequences of AI adoption in the economy (Acemoglu et al., 2022). Despite fairness concerns, algorithms are widely applied across government and industry (Goldfarb and Tucker, 2019; Obermeyer et al., 2019; Aridor et al., 2022). This literature finds that algorithmic decision-making has unanticipated consequences due to misaligned values of platforms, users and regulators (Björkegren et al., 2020; Kasy, 2024). I show that consumers positively value personalization algorithms, and yet seek out the content that aligns with their pre-existing biases. This means feed-diversification is not only costly for platforms, but it is also unlikely to be effective in changing user behavior. These findings are relevant as an increasing number of individuals interact with systems that automatically generate feeds using machine learning algorithms in under-regulated contexts.

2 Background and Data

2.1 The Harms of Social Media and AI

With an average user spending 151 minutes daily on social media platforms, increased social media usage has tightened the scrutiny of the harms these platforms can potentially cause (GWI, 2023). Recent work also shows that social media can adversely affect users' mental health and can encourage the spread of misinformation (Allcott and Gentzkow, 2017; Allcott et al., 2022; Braghieri et al., 2022). Content recommendation algorithms are often accused of boosting engagement with misinformation and hate speech by pushing users into echo chambers of radicalizing content (Pariser, 2011). On the other hand, Gentzkow and Shapiro (2011) show that the internet has reduced the cost of accessing diverse viewpoints, and new communication technologies can also have remarkable societal benefits (Manacorda and Tesei, 2020; Gonzalez and Maffioli, 2024).

However, the link between social media and violence due to misinformation is widely discussed, especially in the case of the Capitol Hill riots on January 6, 2021 in Washington DC, as well as widespread violence in Myanmar and Ethiopia (Narayanan and Kapoor, 2024). Academic research has also linked exposure to radicalizing content on social media

to hate-crimes and politically motivated violence. Müller and Schwarz (2021, 2023) provide evidence that exogenous reductions in social media usage led to a decrease in anti-immigrant hate crimes. This makes the study of social media platforms important as policymakers discuss optimal regulation that prevents the spread of misinformation and hate speech, while preserving users' right to free speech. Furthermore, India's current political context is especially relevant for this study, as the country has seen a rise in hate crimes against minority communities, while increasing internet penetration has made social media a primary source of (mis)information.

2.2 Social Media in India

More than 600 million people in India use social media platforms (GWI, 2023). This makes it one of the largest markets for online platforms in the world, second only to China. On average, 40.2% of the Indian population uses social media, and 67.5% of internet users have used at least one social networking platform (GWI, 2024). Therefore, social media usage is a significant part of the daily lives of a large number of Indians, who spend 141.6 minutes on various platforms every day.[6] These platforms have the potential to amplify marginal voices, but may also have grave consequences for minority communities (Waldron, 2009). With a mobile phone penetration rate of 83%, social media users primarily use handheld devices to access the internet. This is not surprising as India has added more than 500 million mobile broadband connections in the last six years (Waghmare, 2024). While the Demographic and Health Survey (NFHS-5) reports that vulnerable populations have lower access to mobile phones and internet (IIPS and ICF, 2021), social media has become the most widely used platform for public discourse in India, and has been used to communicate underrepresented views (Thakur, 2020).

Social media usage has proven to be harmful in the Indian setting as it has been linked to instances of violence, in the form of mob lynchings, riots, and hate crimes (Banaji et al., 2019). Threats to minority communities, stemming from social media usage in India, are speculated to be bolstered by content recommendation algorithms, which are customization algorithms that employ machine learning technologies. This is because in optimizing content engagement, social media is predicted to generate political filter bubbles or echo chambers (Barberá et al., 2015). Such echo chambers are likely to increase user exposure to more extreme and polarized view points, possibly leading to radicalization (Huszár et al., 2022).[7]

[6]Compilation of statistics retrieved from `https://www.forbes.com/advisor/in/business/social-media-statistics/` on September 27, 2024.

[7]Facebook whistleblower, Frances Haugen, has alleged that the company's personalization algorithms promote extreme content (Haugen, 2021). She also leaked the company's internal documents to show that

India's regulatory framework has struggled to keep pace with the rapid proliferation of social media, leaving significant gaps in addressing the spread of harmful content and its consequences. Despite the introduction of the Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules in 2021, enforcement has been inconsistent, and the regulatory mechanisms lack the teeth to effectively curb the influence of recommendation algorithms that amplify harmful content.[8] If anything, this regulatory framework has been used to stifle dissenting voices, which is in direct contravention of the Santa Clara Principles on Transparency and Accountability in Content Moderation.[9]

Consequently, there is an urgent need for comprehensive policy interventions to tackle the challenges posed by personalization algorithms. However, it is unclear if the continued use of recommendation algorithms is the primary driver of user engagement with extreme content, or if user preferences play a bigger role. This has important implications for the design of regulatory measures in the future, which need to take a multi-pronged approach because behavioral responses may dampen the societal benefits of such policies.

2.3 The Platform

I partner with SM, a prominent social media platform in India, to understand the effects of exposure to extreme content via recommendation algorithms. I study how the nature of online interactions changes with my intervention in SM's rich online social network, comprising almost 200 million monthly active users. SM's user interface resembles that of TikTok, and the platform made massive gains in market share when TikTok was banned in India due to escalating geo-political tensions with China in 2020.

SM is a content-based social network, meaning that users interact with content rather than with other users, unlike X (formerly, Twitter), where users engage with users they 'follow,' and unlike Facebook, where users engage with content from 'Friends,' or from the 'Groups' they join. Connectedness with other users is of little consequence, as is evidenced by the distribution of the number of accounts a user follows in Figure D.4. SM posts, comprising image and video-based posts, are created by influencers on this platform, as most users do not create content themselves. The intervention does not affect the aggregate supply of content because only 1% of SM's users were randomly allocated to the treatment group. Therefore, the intervention left the incentives of these star content creators unaffected.

the company is aware of the harms that algorithms have caused, not just in the US, but also in India. See documents on internally conducted experiments, providing concrete evidence of the problem in India `https://www.nytimes.com/2021/10/23/technology/facebook-india-misinformation.html`

[8] `https://www.freelaw.in/legalarticles/Regulation-of-Social-Media-Platforms-in-India-`. Accessed on September 29, 2024.

[9] See Foundational Principles `https://santaclaraprinciples.org/`. Accessed on January 15, 2025.

Typically, SM users are exposed to a ranked list of posts on their feeds. Here, the ranking was determined by user behavior revealed to the algorithm in previous engagements. Users can scroll over content in the form of short videos, images, and text posts. Due to the new (TikTok-like) features this platform offers, and its multi-lingual interface, SM attracts a large proportion of voters among the urban and rural poor in India. This makes such analysis especially important as little is known about political behavior of this demographic in India or about the users of this massive platform (Aridor et al., 2024).

2.4 Administrative Data

2.4.1 User-Post-Level Data

The administrative data provides information on each post that is viewed or engaged with (by way of sharing or liking) by any given user. The precise time of exposure and engagement is also recorded in the data, which helps identify distinct patterns in usage according to time of the day or day of the week. This allows me to trace the posts a user was exposed to, whether the user chose to engage with the post or not, and under what conditions the posts were engaged with. The user-post level data is used to identify the effects of the intervention on user engagement with posts.

2.4.2 User-Level Data

I observe user characteristics, like their location, gender, age, date of account creation and language in the administrative data. These static user characteristics, along with users' exposure to and engagement with different types of content during the baseline period allow me to analyze heterogeneous treatment effects. The variables and dimensions of heterogeneity used in this analysis were pre-registered with the AEA RCT Registry (Kalra, 2023).

I provide a descriptive summary of user characteristics, as well as engagement at baseline in Table 1. This Table also verifies the balance of observable user characteristics across the treatment and control groups.

2.4.3 Post-Level Data

The platform characterizes posts by broad tag genres, based on user generated hashtags.[10] Further, the text on the images/ videos in the user generated posts is a rich source of information. I adopt various methods to analyze the text data, starting from LDA topic modeling to understand the broad themes in the posts, to sentiment analysis to understand

[10]I do not have access to the algorithms that allocate posts to these genres or categories.

the tone of the posts (Ash and Hansen, 2023). This helps in measuring political slant of more than 20 million posts that users engaged with, during the course of the experiment.

The focus of this paper is on political posts that target India's sizeable minority communities. Therefore, I repeat the analysis of the text data on the subset of posts that are in the Politics or Devotion/ Religion genre. The descriptive analysis of the text, detailed in Appendix I, highlight the need for contextual embeddings that accurately characterize the potential harm that a post may cause.

2.5 Toxicity Classification

The administrative data provides user-post level data on viewership and engagement. To measure the main outcome variable, i.e. toxicity of shared posts, I further process the text from images in the post data (using OCR), to classify them as toxic. I problematize posts that are a direct threat to the safety of a group or individual, but also disrespectful posts that are likely to make one leave a discussion.

Specifically, I use Perspective's machine learning algorithms, developed by Jigsaw at Google, to identify toxicity in the Hindi text extracted from about 20 million posts.[11] Perspective offers functionality in various languages, including Hindi, and is therefore able to preserve the context of the text in the classification process which could potentially be lost in a translation to English. Toxic content is defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion."

Perspective is a widely recognized machine learning solution for toxicity detection. It leverages transformer-based deep learning models, trained on millions of comments annotated by multiple human raters who evaluate contributions on a scale ranging from "very toxic" to "very healthy" (Fortuna et al., 2020). Transformer-based models, like GPT and BERT, use self-attention mechanisms to process entire sentences at once, enabling them to capture long-range dependencies in text. This makes these models sensitive to context (Vaswani, 2017) Perspective's Machine Learning models are being widely adopted to identify and filter out abusive comments on platforms like New York Times, and are also being frequently used in academic research (Jiménez Durán, 2022).

I construct a binary variable, labelled "toxic," which takes value 1 when Perspective's toxicity score on a post is higher than 0.2. The 0.2 threshold is chosen to maximize the criterion of true positive rate in the classification and F1 score, in particular.[12] In Figure

[11] Jigsaw is a research unit within Google that builds technology to address global security challenges. For more information, see `https://jigsaw.google.com/`.

[12] The F1 score is a metric used to evaluate the performance of a model, particularly in tasks like classification. It is the harmonic mean of precision (how many of the predicted positive results are actually correct) and recall (how many of the actual positive results the model successfully identified).

I.2, I show that 0.2 satisfies this threshold selection criterion, where the true labels for a random set of posts in the confusion matrix were determined by human raters, who were Hindi-reading undergraduate students at Brown University. While a 0.1 threshold has a higher rate of correctly classifying toxic speech, the F1 score is maximized at 0.2 and is a more appropriate measure, especially when there is imbalance in the dataset (Dell, 2024).

I further validate this threshold in Appendix I by comparing the performance of this cut-off with other methods of detecting harmful content, and providing examples of Hindi posts from the platform along with the continuous toxicity scores. This paper focuses on toxic posts that target India's minority Muslim population. Since Perspective is also correctly classifies homophobic and sexist content as toxic (for instance), I replicate the analysis on a subset of political and religious posts, where a majority of the toxic posts are anti-Muslim. Engagement with such content has been shown to create positive affect among social media users (Schmid et al., 2024), which is likely to increase addiction to social networking sites (Jo and Baek, 2023).

2.6 Survey Data

I supplement the outcome measures on platform usage, that are available in the administrative data, with an online survery that was sent out in three waves between May 2023 and July 2024. The protocol involved sending out a survey to users' registered WhatsApp numbers through the platform's WhatsApp business account. This received a low response rate, despite the survey being heavily incentivized. The second protocol, in the post period of the experiment, was a telephonic survey with a higher response rate.

In this paper, the survey data is used only to supplement the main results from the administrative data in Section 4. These data were especially useful in understanding how treated users spend their time if the intervention caused them to disengage from SM. For instance, the survey asked users about their time spent on other social media platforms, and their attitudes towards redistribution.

3 Experimental Design

3.1 Sample

I collaborated with SM to design and conduct a large-scale, long-term experiment that exposed users to content randomly drawn from the corpus of 2 million posts generated in the Hindi language each day. Out of the 200 million users on the platform, about 2 million users were treated at the start of the experiment in February, 2023. Approximately 6 million

were selected to be in the control group to prevent contamination due to other experiments running on the platform.

I limit the analysis to Hindi language users, who constitute about 45% of the total user base, to ensure the accuracy of the text analysis, given my native proficiency in Hindi. Further, I only include active users in the sample, defined as individuals who viewed at least 200 posts during the baseline period (5% of the experimental sample). These cuts reduce the sample size to 231,814 users, with 63,041 in the treatment group and 168,773 in the control group. Table 1 shows that 70% of users in the sample are men. This figure aligns with the gender distribution of social media users in India as reported by NFHS-5, which found that 41% of women in India do not use the internet (IIPS and ICF, 2021). Further, the average user in the sample created their account in 2022, and almost two years after TikTok was banned in India. This is consistent with the estimated growth in internet penetration in India around 2022, when the proportion of internet users increased from 20% in 2018 to 46% (World Bank, 2022).

Users with a higher proclivity to toxic content are among the oldest users on the platform (Figure D.16), and the average user spent close to 7 hours on the platform during the baseline period of 31 days (Table E.1). This is lower than the estimated time spent on social media in India (141.6 minutes) each day, indicating that the average user was also actively consuming content from other social media platforms (GWI, 2024). The platform's integration with WhatsApp, a unique feature, indicates that its users are also very active on WhatsApp, the most popular social networking application in the country (GWI, 2023). As a result, the platform is representative of internet users in India who largely use SM to consume content that is suitable for sharing as WhatsApp forwards in private conversations. This also has implications for the type of content SM users consume. For example, the average user in my sample viewed 1,087 posts during the baseline period, with most this content falling under the categories of "greetings" and "devotion" in Figure D.16.[13]

3.2 Randomization

Treatment was randomly assigned at the user level to 1% of SM's user base, which includes both active and inactive users. User IDs were picked randomly at the start of the experiment, and selected users were assigned to the treatment for the entire duration of the intervention. Similarly, control users were also selected at the start of the intervention to ensure that

[13]The "greetings" genre includes posts that wish users good morning. This is a peculiar use of social networking platforms in India, which has received some attention. See, for example, `https://www.wsj.com/articles/the-internet-is-filling-up-because-indians-are-sending-millions-of-good-morning-texts-1516640068`.

their outcomes were not subject to contamination due to other AB tests/RCTs running on the platform. Since users must opt into being randomly assigned to treatments for market research and AB tests when they create their account, they were unaware of their participation in the experiment, reducing the likelihood of selection bias. This is worth emphasizing as previous work on social media algorithms may not be generalizable outside of lab-like settings, where users are aware of the experiment and may change their behavior accordingly (Guess et al., 2023b).

I verify the validity of randomization in the treatment assignment across the sample by assessing balance in observable user characteristics across the treatment and control groups. I also consider various user attributes, including gender, state and city of residence, and the week in which a user first created their account, as well as various measures of baseline usage, such as the total number and proportion of posts viewed.

I cannot reject the hypothesis that the treatment assignment was uncorrelated with user characteristics, either individually or jointly. Table 1 provides estimates for a randomly selected set of attributes, showing balance in behavior at baseline with respect to viewing and sharing all types of posts, including toxic posts, across treated and control users.

3.3 Control

Users in the control group continued to receive these standard algorithmic recommendations, while I intervene on the personalization algorithm for treated users. SM's personalization algorithm customizes user feeds based on their engagement history, using a Field Aware Factorization Machines (FFM) algorithm (Aggarwal et al., 2016). This algorithm generates a vector of preference weights for each user with respect to various post attributes, which are calculated using matrix factorization methods.[14] These vector weights in the space of certain post features are referred to as embeddings in the machine learning/ deep learning literature (Athey and Imbens, 2019; Dell, 2024). This generates a ranking of posts for each user, and new posts are recommended daily according to this order.

The personalization algorithm generates these preference weights, or embeddings vectors, based on various (latent) features, which might represent a user's or post's affinity for characteristics such as humorous or toxic content, for instance. In general, these features are not interpretable but are learned by the algorithm to maximize user engagement with the platform. Appendix H provides a general overview of how these algorithms function, using data on recent user-post engagements and a simple example. This example helps fix ideas

[14] Although the embedding vectors for the treatment and control groups are determined simultaneously, the intervention does not spill over to the control group because the embeddings are replaced for only 0.5% of SM users.

and simulate the general properties of the personalization algorithm.

3.4 Treatment

Treated users were shown posts that were not ranked according to user preferences but randomly drawn from the entire corpus of content in their chosen language. This effectively randomized the probability of a post being recommended to a user.[15] Thus, while the control group's recommendations were personalized according to their past behavior, the treatment group experienced post recommendations based on random assignment.[16] FiguresD.1 shows that treated user feeds changed along a variety of dimensions characterized by different topics. However, Figure D.2 shows that users with high interest in toxic content witnessed the largest decreases in exposure to such content, and is also the policy-relevant dimension. Therefore, this intervention is a blunt instrument for a precision job as targeting specific types of problematic content is not feasible for the platform.

I demonstrate the key properties of the content distribution in the treatment group by simulating a simple recommender system that generates the probability of the personalization algorithm assigning a post to a user (see Appendix H for details). First, the distribution of content assignment probabilities among treated users approximates a normal distribution centered around the average probability observed in the control group (Figure H.3). This is predicted by the Law of Large Numbers (LLN), as the assignment probabilities are randomly picked from the set of control probabilities for treated users on each day of the intervention period. Second, the LLN also predicts a smaller spread in assignment probabilities for the treatment group given that the variance of these probabilities is divided by the number of control group users.

Crucially, the treatment has a greater effect on users with more extreme preferences regarding the toxicity of content. Figure H.4 shows that users with preferences closer to the average did not experience large differences in their assignment probabilities, and therefore the content feed, when treated. This important characteristic of content distribution is formally discussed as treatment intensity in Section 4. Contrary to expectations, the assign-

[15]The random draw of posts for treated users were generated by replacing the algorithmically generated embedding vectors with randomly picked multidimensional embeddings for each treated user. That is, for each treated user, the vector of preference weights is just a random draw of numbers. See Appendix H.

[16]The "random embeddings" for treated users were uniformly sampled from an epsilon ball whose centroid was given by the mean embedding in the control group and the radius was twice the sum of variances in that vector. In particular, $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i$, $\sigma^2 = \frac{1}{\nu}\sum_{i=1}^{\nu}(\boldsymbol{x}_i - \boldsymbol{\mu})^2$, where $\boldsymbol{x}_i$ represents the embedding with bias for user i and ν is the total number of users. Formally, for each user embedding, the "random algorithm" uniformly sampled a point from an epsilon ball with the centroid $\boldsymbol{\mu}$ as the center and radius $2 \times$ variance of control embeddings. Then, $\boldsymbol{\rho} \sim \mathcal{U}(\text{Ball}(\boldsymbol{\mu}, 2\sigma^2))$, where $\boldsymbol{\rho}$ is the newly sampled embedding for the user and $\mathcal{U}(\text{Ball}(\boldsymbol{\mu}, 2\sigma^2))$ represents a uniform distribution within the epsilon ball centered at $\boldsymbol{\mu}$ with radius $2\sigma^2$.

ment of average preference weights to treated users does not bias their exposure to popular content, as seen in Table H.1.

The experiment is policy relevant as social media platforms like SM often introduce some randomly drawn posts in personalized feeds to expose users to a more diverse set of content. SM typically randomizes a small proportion of posts in a user's feed to maximize learning in an algorithm operating on an exploration-exploitation frontier. While this helps the platform to continuously learn about user preferences, I show that it also diversifies the content that users consume (Kleinberg et al., 2022). The intervention began on February 10, 2023 and continued until the end of the year. Administrative and survey data on relevant outcomes were gathered for the baseline period (December 2022), the intervention period (February to December 2023), and the post-intervention period (January to March 2024).

3.5 Descriptive Statistics

The algorithm keeps users engaged online. Users value the personalization algorithm because it decreases the cost of searching for preferred content. By making content discovery more difficult, the treatment reduces overall engagement with the platform, as reflected in the total number of posts viewed and shared (Table 2). This suggests that users gain value from the algorithm and disengage from the platform when it is turned off.

Table E.1 shows disengagement in all aggregate measures of platform usage. There are negative and statistically significant treatment effects on the number of logins per month but positive effects on the probability of leaving the platform. The average treated user reduced their total time on the platform by 2.5 hours compared to the average control user, who spent close to 7 hours per month.

These results suggest that the intervention was costly for SM, as the platform generates revenue from the time users spend on the app and their attention to advertisements. In particular, Table 2 shows that on average, treated users viewed 35 fewer posts, while control users viewed about 250 posts. Back-of-the-envelope calculations suggest that if the intervention is upscaled to the entire platform, SM loses $45,817 in advertising revenue in the first month of the intervention.[17] This is less that 1% of SM's reported revenues in 2023.

Treated users view less toxic content. The direction of the treatment effect on the number of toxic posts viewed, or the treatment intensity for the average user, is unclear a priori, as it is expected to be positive for users who do not prefer toxic content and negative

[17]The estimate was obtained using the price of INR 0.55 that SM charges advertisers per 1,000 impressions, in an educational ad campaign that I designed. It was then converted to USD using the exchange rate of 1 USD = 84.03 INR, as of October 6, 2024.

for those who do.[18] Therefore, the average effect depends on the distribution of user types in the sample as well as the average probability of being assigned toxic content. [19]

Figure D.3 shows that during the first month of the intervention's implementation (i.e., February 10 to March 10, 2023), the treatment group was exposed to fewer toxic posts on average due to the random content delivery. Table E.2 shows that there is an average reduction of 14% in the toxicity of posts viewed, when continuous toxicity score is not dichotomized in column (1).

Online sharing behavior is inelastic The average user viewed fewer toxic posts in the intervention period than in the baseline period. I therefore expect a reduction in the total number of toxic posts shared, as treated users face positive search costs in seeking out toxic posts to share, and exposure to more diverse content may change their attitudes toward toxicity. However, I find that while treatment reduced the proportion of toxic posts viewed by 8.7%, it increased the proportion of toxic posts shared by 7.8% (Table 2). This is because even though the average treated user saw less toxic posts and shared a smaller number of toxic posts, they shared far fewer posts of other categories.

Furthermore, the decrease in the number of toxic shares (20%) is not as large as the decrease in the number of toxic views (27%). Therefore, the elasticity of toxic sharing with respect to toxic viewing—defined as the ratio of the percentage change in the number of toxic posts shared to the percentage change in the number of toxic posts viewed—is less than 1. I reject the null hypothesis that users behave mechanically with respect to the toxic content they are exposed to (or that the elasticity of toxic sharing with respect to toxic viewing equals 1) with a p-value of 0.002.

Inelasticity in user behavior is corroborated by survey evidence in Figure D.18 that shows no differences in political attitudes of treated and control users despite the treated users being exposed to a random draw of posts. These outcomes are precisely estimated meaning that small differences in attitudes can also be ruled out as I surveyed more than 8,000 users, about half of whom were treated.

Treatment induces behavioral responses to seek out content. On average, the treatment effect on the ratio of toxic shares to toxic views is positive. Thus, the average treated user changed their behavior in response to the intervention, sharing toxic content at a

[18]To build intuition, this can be observed in the simulated recommendation algorithms for the treatment and control groups in Figure H.4, where all treated users are exposed to toxic content with a uniform probability.

[19]The probability of being assigned toxic content does not necessarily equal the inverse of the number of toxic posts in the corpus, as it is determined by the cross-product of user and post embeddings (detailed in Appendix H). As a result, the average treatment intensity does not equal zero in this experiment.

higher rate and offsetting the negative treatment effect on the number of toxic posts shared. To illustrate this, the treatment effect on the number of toxic posts shared is decomposed as follows:

$$\text{Toxic Shares} = \frac{\text{Toxic Views}}{\text{Posts Viewed}} \cdot \text{Posts Shared} \cdot \frac{\text{Proportion Toxic Shares}}{\text{Proportion Toxic Views}}, \tag{1}$$

where the first term in the decomposition corresponds to the mechanical change in exposure to toxic posts due to the intervention, the second term corresponds to the disengagement effect that reduced platform usage, and the third term corresponds to the change in behavior upon viewing diverse content.[20]

On average, I find that the exposure and disengagement effects contributed to 66% of the reduction in the number of toxic posts shared in this empirical decomposition. This suggests that the change in behavior, as seen in the ratio of toxic shares to toxic views (the residual 34%), plays a significant role in dampening the aggregate effect of the intervention. This is because if the behavior change in the ratio had been less than or equal to zero, the treatment effect on the number of toxic posts shared would have been more negative, yielding greater benefits for society.

Treated users search more. Treated users were more likely to use the search feature on the platform (Table E.1), which complements the evidence on the stickiness of sharing behavior, as my measure of shares includes posts accessed through both the trending feed and search tabs. This finding also aligns with the fact that treated users were more likely to view fewer posts during the intervention period, as my measure of views excludes posts accessed through the search tab.

While searching offers an intuitive channel for the positive effect on the ratio of toxic posts shared to toxic posts viewed, it is less likely to be driving these effects, as searched posts constitute 0.01% of viewed posts and 0.004% of shared posts.

An Illustration of User Behavior. While the treatment significantly decreased the number of toxic posts viewed by the average user, the reduction in the number of toxic posts shared was not as large. This suggests, as supported by the descriptive evidence, that user behavior is not malleable or elastic, as the change in sharing behavior did not correspond proportionally to the change in views.

For example, consider a user who is served 15 posts in a day, of which they share 9. If they are served 5 toxic and 10 non-toxic posts and share 2 toxic and 7 non-toxic posts,

[20]The term *mechanical* is misleading because some of the change in exposure is endogenous, as users can change the total number of posts viewed in response to the intervention.

then the proportion of toxic posts shared is $\frac{2}{9} \times 100 = 22\%$. Now, consider a treated user who views 2 toxic posts and 7 non-toxic posts, and suppose they share 1 toxic post and 3 non-toxic posts. Thus, they are disengaged from the platform, sharing a total of 4 posts instead of the 9 they would have shared if they had not been treated. Note that they also view a smaller number of posts.

This example illustrates that even though the average user views and shares fewer toxic posts upon being treated, the proportion of toxic shares increase. This occurs because 1 out of 4 shares is toxic under treatment, meaning the proportion of toxic shares is $25\% \geq 22\%$.

The stickiness in user behavior is also reflected in the ratio of toxic shares to toxic views. For example, control users shared 2 out of 5 toxic posts they viewed (40%), while treated users shared 1 out of 2 (50%). Another useful statistic for building intuition is the elasticity of toxic shares with respect to toxic views. Due to the intervention, the percentage change in toxic shares is -50%, while the percentage change in toxic views is -60%. Thus, elasticity of toxic shares with respect to toxic views is 0.83.

4 Results: Four Facts

In this section, I present four key findings from the empirical analysis of the experiment, followed by a discussion of their broader implications.

Fact I: Disabling the algorithm has heterogeneous effects

The intervention assigned average content from the full library of posts, reducing the amount of particular content for users who typically consumed above-average amounts and increasing it for those with below-average consumption.[21] Therefore, treatment intensity is higher for extreme users, as their baseline exposure to toxic content differed more significantly from the average feed.

To examine potential heterogeneous effects, I rank users based on the percentage of toxic content in their feed at baseline. This approach allows me to compute the effects on users with low, medium, and high levels of baseline toxic exposure. Using baseline exposure to represent user types is accurate, as the personalization algorithm recommends posts based on users' past behavior. Figure 1 shows that the treatment effect on the proportion of toxic views is negative for users with a high degree of toxic exposure at baseline (Q3–Q5) and is positive for those with low baseline exposure (Q1 and Q2).

[21] This highlights the importance of targeting in my setting, as the treatment is different for different individuals as in Duflo et al. (2011)

Fact II: Usage declines when personalization is turned off

Personalization likely makes it easier for users to find the content they prefer, especially for those with more extreme preferences, as average content is further from their favored material. When personalization is removed, as shown in Figure 2, platform usage changes based on users' baseline toxic exposure. Users with the highest degree of baseline toxic exposure (Q5) reduced the number of posts viewed the most, by 23.2%. Views also declined for users in the middle of the distribution (Q2–Q4). However, those with the lowest toxic exposure at baseline (Q1) did not reduce their views, even though the treatment changed their feed more than it did for those in the middle.

Table E.1 shows that treated users were 20% more likely to leave the platform after the first month of the intervention compared to control users. However, this effect is not heterogeneous by baseline exposure (Figure D.11), suggesting that Q5 users largely disengaged on the intensive margin rather than on the extensive margin.

This may raise concerns about differential attrition across the treatment and control groups. Specifically, the main estimates may be biased if the type of treated users who left the platform differed from those who stayed. However, controlling for baseline engagement with toxic content and treatment status, Table E.3 shows that leavers and stayers were balanced on observable characteristics. In Appendix F, I show that the Lee bounds for the main outcomes are tightly estimated as the magnitude of attrition is small (Lee, 2009). This confirms that the analysis by user type is robust to differential attrition.

Fact III: Sharing of toxic content is nearly inelastic but more elastic for users initially exposed to more toxicity

While viewing posts on the platform can be a more passive decision, sharing posts with others is an active choice. This distinction is reflected in the data, as Figure 3 shows that Q5 users, who had higher baseline exposure to toxic content, reduced the number of toxic posts they shared when personalization was removed. In contrast, Q1 users, with the lowest baseline exposure to toxic content, did not increase the number of toxic posts they shared, even though they were exposed to a substantial increase in toxic posts.

To assess how users' decision to share was affected by exposure, I define the elasticity of toxic sharing as the ratio of the percentage change in toxic posts shared to the percentage change in toxic posts viewed. If the elasticity is 1, then decreasing the toxic content a user is exposed to reduces the amount of toxic content they share by the same proportion. If the elasticity is below 1, users decrease their sharing by less than their exposure, and if it is zero, sharing does not change with exposure.

I find that the elasticity is below 1 for all users, ranging from 0.08 for Q1 users to 0.69 for Q5 users, with the highest elasticity observed for those who viewed the most toxic content at baseline. This finding helps explain why Q1 users did not experience a decrease in the total number of posts viewed during the intervention period, as was expected for extreme users, despite the high and positive treatment intensity in Figure 2.

Figure D.12 shows that while treated users were more likely to use the platform's text search feature, the treatment effect on the number of times a user searched for any content (per post viewed on the landing page) is not heterogeneous. While illuminating, this evidence is suggestive as searched posts constitute 0.01% of viewed posts and 0.004% of shared posts, as discussed in Section 3.5.

Fact IV: Behavioral responses dampen benefits of regulations

The intervention makes it more difficult for treated users to discover content that would have been recommended by the personalization algorithm, which may lead them to share fewer posts of the type they are inclined toward, assuming they do not change their behavior. Alternatively, if users do change their behavior, they may seek out their preferred content and share a higher proportion of it.

Figure 3 shows that Q5 users saw the largest decrease in the number of toxic posts shared upon treatment, with the treatment effect monotonically decreasing across user types. However, relative to the reduction in toxic posts viewed, these users actually saw an increase in the proportion of toxic posts shared.

This increase in the proportion of toxic posts shared by Q5 users suggests a behavioral change, as they amplify their toxic behavior by sharing a higher proportion of toxic posts they view. Figure D.7 decomposes the effect on the number of toxic posts shared, based on the empirical decomposition in Equation 1 for different user types. The results show that for Q5 users, the increase in the ratio of toxic shares to toxic views contributes to 39% of the total effect on the number of toxic posts shared. If this effect had been zero, the total effect would have been more negative, especially for Q5 users. In other words, the behavioral response in the ratio of toxic posts shared to toxic posts viewed dampens the societal benefits of the intervention, which are driven by the decrease in the total number of toxic posts shared.

5 Model

In this section, I introduce a model to rationalize opposing effects of the intervention on the absolute and relative quantities of toxic shares. I evaluate counterfactual policies that target toxic content on social media as such policies cannot be implemented on the field.

5.1 Setup

I model a strategic interaction between the platform's algorithm and each user. The players' objective functions and strategies are described below.

5.1.1 Platform and Algorithm

For each user i, the platform maximizes N_i, the total number of posts i views, assumed to be a continuous variable.[22] The algorithm assigns toxic and non-toxic posts with probabilities q_i^t and $(1 - q_i^t)$, respectively, to maximize engagement.[23] Then, N_i is endogenously determined by the user for these given assignment probabilities chosen by the algorithm, as shown next in the users' optimization problem.

5.1.2 User

Consider a social media user i, who picks the utility maximizing number of posts to view, $N_i = N_i^t + N_i^n$, for given assignment probability q_i^t determined by the algorithm because N_i^t is determined as $N_i q_i^t$, and N_i^n as $N_i(1 - q_i^t)$. She also chooses S_i^t (toxic) and S_i^n (non-toxic) posts to share out of the total N_i^t and N_i^n posts viewed, in order to maximize consumption value from both viewing and sharing posts. This determines the continuous interval of posts shared $[0, S_i]$ from $[0, N_i]$ interval of posts viewed, where $S_i = S_i^t + S_i^n$. Effectively, the user chooses the proportion of posts shared that are toxic as $s_i^t = S_i^t / S_i$.

One motivation for this model is that social media users identify with some social categories that can represent users gender, caste, or religious affiliations, for example. As in Akerlof and Kranton (2000), I introduce the set of categories C based on some observable group characteristics or social identities. In this setting, I characterize the identities with user proclivity for toxic context, so that user type is given by baseline exposure to toxic

[22] I follow Becker (1991) in adopting the continuity assumption.

[23] The platform's problem is a simplification of the actual problem faced by social media platforms, where the platform also optimizes the number of likes, shares, comments, number of ads shown to each user, and the price of advertising. The rank of a post on the content feed is now reduced to a single number, i.e. the assignment probability. I abstract away from the exact process that translates views to advertising revenues as the objective is to mimic the incentives of a simple algorithm in order to analyze user responses, and not the algorithm itself.

content. That is, $C = \cup_{k=1}^{5} Q_k$, where Q_k represent quantiles in the distribution of users' exposure to toxic content at baseline in Section 4. User preferences for toxic content are denoted by $p_i^t \in [0,1]$ in the model, and are represented by users' exposure to toxic content at baseline in the empirical analysis. Then, each user i belongs to some category $c_i = c(p_i^t)$, so that she has a conception of own categories or types, and those of other people.

Following Butera et al. (2022), users also derive public recognition utility from sharing posts that reflect conformity with perceived social norms followed in their social group $c(p_i^t)$. I assume that they learn about other users' tastes for toxic content from the content feeds, consistent with studies in media psychology that show users rely on algorithmic curation to learn about appropriate discourse from other users' behavior (Masur et al., 2021). That is, users are able to infer their social category's tastes from the content they view on the platform. This implies that q_i^t also depends on c_i. The user's objective is to maximize utility she derives from viewing and sharing posts,

$$\begin{aligned}\max_{s_i^t, S_i, N_i} u(s_i^t, S_i, N_i; q_i^t, c_i) = & \underbrace{\beta N_i - \alpha(N_i - S_i)^2 - \eta S_i^2}_{\text{consumption utility}} \\ & - \delta S \quad \underbrace{(1-\theta)\left(\log\left(\frac{s_i^t}{p_i^t}\right)\right)^2}_{\text{disutility of deviating from own tastes}} \\ & - \delta S \quad \underbrace{\theta\left(\log\left(\frac{s_i^t}{q_i^t(c_i)}\right)\right)^2}_{\text{disutility of deviating from category's tastes}}\end{aligned}$$

where, β is the weight assigned to consumption utility received from viewing posts, N_i. A user is assumed to incur some disutility if an additional post she views is not shareable according to $\alpha(N_i - S_i)^2$. $\eta > 0$ is the cost of increasing total number of posts shared. Users have self-image concerns which are modeled as disutility from sharing toxic content in a way that deviates from users' own preferences, p_i^t. This disutility is parameterized with $(1-\theta)$, and δS normalizes the utility function in terms of number of posts. I use the log form in the model because it is shown to fit the experimental data on toxic shares well. However, all model predictions are shown to be robust to functional form choices in Appendix A.

I assume that $\theta \in [0,1]$ is the weight users put on their perception of society's tastes for toxic content so that users update their behavior in line with their perception of norms at some rate θ. This parameter measures the "influence" on account of exposure to the algorithmically generated feed. The disutility from sharing toxic content depends on how users' sharing behavior differs from a reference level which is given by a combination of what

users' think that others do, $q_i^t(c_i)$, as well as their own tastes, p_i^t (DellaVigna et al., 2012).

5.2 Dynamics and Equilibrium

The timing of the strategic interaction is as follows. The algorithm chooses assignment probabilities $q_i^t(c_i)$ to maximize engagement. Next, a user decides the total number of posts to view N_i, or the time she spends on the platform upon observing the assignment probabilities. Then, the user chooses the total number of posts to share, S_i, which offers the user with some consumption utility. Finally, she chooses the fraction of shared posts that are toxic, $s_i^t = S_i^t/S_i$, for given exposure and sharing decision.

Consider two time periods, so that $\tau = 0$ and $\tau = 1$ represent the baseline and the intervention periods, respectively. In each time period, the game unfolds according to the timing described above. The algorithm and users are in equilibrium at baseline $\tau = 0$, irrespective of treatment status.

The model allows a characterization of user preferences in terms of baseline exposure to toxic content due to the platform's profit-maximization problem. This is because the algorithm's first order condition implies that the equilibrium assignment probabilities exactly equal the respective users' tastes for such content.

Lemma 1. *The platform's engagement maximization problem implies*

$$q_{i,\tau}^t(c_i) = p_i^t \tag{2}$$

That is, the algorithm assigns toxic posts with probability equal to user's intrinsic tastes for toxic content.

This result enables the characterization of user preferences as $q_{i,0}^t(c_i) = p_i^t$ for all i in equilibrium at baseline ($\tau = 0$). The algorithm also maps individual preferences to their respective social categories c_i as the categories are defined in terms of preferences p_i^t. All proofs are contained in Appendix B.

The equilibrium conditions build up to the estimation strategy for the main parameter of interest, θ. This follows from the first order condition of the consumer's utility maximization problem which gives the optimal sharing function in equilibrium.

Lemma 2. *For a utility maximizing agent i,*

$$s_{i,\tau}^t = \left(q_{i,\tau}^t(c_i)\right)^{\theta} \left(p_i^t\right)^{1-\theta} \tag{3}$$

That is, users place a weight of θ on perceived social norms while choosing the proportion of posts shared that are toxic.

Therefore, the content feed informs the user's perception of her category's tastes for toxic content which affects sharing behavior because of exposure to the content feed at the rate of θ. This is tenable as users know that the content feeds are tailored according to preferences of all users in the same social group c_i because $q^t_{i,\tau}(c_i) = p^t_i$ in equilibrium.

5.3 Model Predictions

The control algorithm assigns toxic posts to each user i at time τ with probability $q^t_{i,\tau}$. However, in the treatment group, the probability of being assigned toxic content is picked uniformly at random, each day during the intervention period, from the set of all possible assignment probabilities in the control group, that is $\bar{q}^t = \mathrm{E}[q^t_{i,\tau}|i \text{ assigned to control group}]$. The model enables an analysis of comparative statics for the exogenous variation in assignment probabilities under treatment.

I consider a targeted policy where users with higher proclivity to toxic content, that is $p^t_i > \bar{q}^t$, are treated with diversified or randomized feeds. This is because the empirical analysis demonstrated that the treatment increased the number of toxic posts viewed by users with the lowest proclivity to toxic content. Such an effect is not desirable, either from the platform's perspective (Q1 users do not see what they like), or that of the planner (Q1 users see more toxic content). Diversifying content feeds for users with $p^t_i > \bar{q}^t$ (Q4 and Q5 users) allows a more direct approach to reducing toxic exposure for the most toxic users.[24]

Proposition 1. *Treatment effect on exposure to toxic content is negative for user i with higher proclivity towards toxic content. Further, this effect is smaller for larger p^t_i.*

That is, the treatment intensity is negative for users with higher exposure to toxic content at baseline, and more negative for users with more extreme preferences. Figure 1 confirms the model's predictions to show that the treatment effect on the number of toxic posts viewed is negative for Q4 and Q5 users. This figure also shows that the treatment intensity is larger (in absolute terms) for Q5 users, compared to Q4 users. Next, the model predicts that the users with higher baseline exposure to toxic content view fewer posts overall.

Proposition 2. *For user i with $\alpha, \beta, \delta, \eta > 0$, $\theta \in [0, 1]$ and $p^t_i > \bar{q}^t > 0$,*

$$\frac{\partial^2 N_{i,\tau}}{\partial p^t_i \partial \bar{q}^t} \geq 0 \tag{4}$$

[24]Such a targeted policy could not be implemented in the field experiment because the platform does not want to target users or posts by toxicity so as to maintain political neutrality.

That is, for marginal increases in the average probability of being assigned toxic content $\bar{q}^t$, users with higher proclivity to toxic content view more posts.

Intuitively, this is because the treatment exogenously lowers the probability of being assigned toxic content to the control mean when $p_i^t > \bar{q}^t$. Therefore, marginal increases from $\bar{q}^t$ bring this probability closer to the user's true taste for toxic content, p^t. In other words, the treatment effect on the total number of posts viewed is expected to be negative and smaller, and so is predicted to have a bigger impact (in absolute terms) on more toxic users.

In Figure 7, I simulate the model's predictions on the total number of posts viewed, under two regimes: treatment and control. The control users continue viewing the optimal number of posts in equilibrium but treated users face a higher treatment intensity, and so choose to view fewer posts in total. Given that the treatment intensity is higher for users with more extreme preferences, the model also predicts that users with higher exposure to toxic content at baseline spend even less time on the platform. Figure 2 shows that I cannot reject this hypothesis because Q4 to Q5 users view fewer posts overall. Further, the reduction in number of posts viewed is larger for Q5 users.

Proposition 3. *For user i with $\alpha, \beta, \delta, \eta > 0$, $\theta \in [0, 1]$ and $p_i^t > \bar{q}^t > 0$,*

$$\frac{\partial^2 s_{i,\tau}^t}{\partial p_i^t \partial \bar{q}^t} \geq 0 \tag{5}$$

That is, marginal increases in the average probability of assigning toxic content leads to larger increases in the proportion of shares that are toxic for users who prefer such content.

For the given rate of influence θ, the model predicts that the decrease in the proportion of toxic posts shared is larger for more toxic users. Figure 7 graphically demonstrates another model prediction that the effect on $s_{i,\tau}^t$ is concave in user tastes. Bringing this prediction to the data, Figure D.13 shows that the treatment effect on proportion of shares that are toxic is negative for users in Q5. Finally, the model predicts that the probability of sharing toxic posts conditional on viewing them is higher among the treated. This is because users with higher proclivity to toxic content are less likely to view toxic content, but are more likely to share it in line with baseline behavior.

Proposition 4. *For user i with $\alpha, \beta, \delta, \eta > 0$, $\theta \in [0, 1]$ and $p_i^t > \bar{q}^t > 0$,*

$$\frac{\partial^2 (s_{i,\tau}^t / \bar{q}^t)}{\partial p_i^t \partial \bar{q}^t} \leq 0 \tag{6}$$

That is, increasing the average probability of assigning toxic content leads to smaller changes in the proportion of toxic shares out of toxic views for users who prefer such content.

Figure 7 predicts that the ratio of toxic shares to toxic views is larger for treated users with higher exposure to toxic content at baseline. The empirical results are in line with this prediction as Figure 4 finds that the average effect on this ratio is driven by Q4 and Q5 users who share toxic content at a higher rate.

5.4 Estimation

5.4.1 Measurement

Algorithmic content assignment is assumed to be equilibrium at baseline with $q^t_{i,0} = p^t_i$, as shown above. This means that user behavior is determined by substituting for $q^t_{i,0}$ into the optimal sharing function (8) as $s^t_{i,0} = p^t_i$. Therefore, preferences are measured using baseline sharing behavior during the intervention period.

Further, I proxy user's probability of being assigned toxic content by the algorithm $q^t_{i,\tau} \equiv v^t_{i,\tau}$, where $v^t_{i,\tau}$ is the proportion of posts viewed that are toxic.[25] As a result, Lemma 2 provides that sharing in equilibrium is a function of sharing at baseline, and the type of posts viewed. I introduce preference shocks into the sharing behavior to get,

$$\log s^t_{i,1} = \theta \log v^t_{i,1} + (1-\theta) \log s^t_{i,0} + \mu w^t_{i,1} \tag{7}$$

where, $w^t_{i,\tau}$ represents iid preference shocks or unobserved heterogeneity in sharing behavior. The main parameter of interest θ is interpreted as the influence of exposure. This is in line with the idea that users expand their view of socially acceptable things to say in public discourse, by observing the content that is recommended to them by the algorithm.

θ cannot be directly estimated through equation (7) as exposure to toxic content is constant for all users i in the treatment group, $v^t_{i,1} = \bar{q}^t$. Further, the preference shocks are interpreted as user "moods" correlated with sharing behavior $s^t_{i,\tau}$. A steady state condition is used to identify θ.

5.4.2 Steady State

This system is in steady state when the probability of viewing and sharing toxic posts $(v^t_{i,\tau}, s^t_{i,\tau})$, are stable over time. The steady state condition is also the identifying condition because in the absence of any exogenous changes to assignment probabilities, user behavior should be the same in each time period. As a result, any changes in the probabilities of sharing toxic content are due to changes in exposure to toxic content. That is, θ is identified

[25] $v^t_{i,\tau}$ can also be some fraction of $q^t_{i,\tau}$ because these variables are assumed to be continuous as $N_{i,\tau}$ and $S_{i,\tau}$ are not discrete in this framework.

when the following assumption is satisfied,

$$\begin{aligned}\log s_{i,0}^t &= \log s_{i,1}^t \\ &= \theta \log v_{i,1}^t + (1-\theta)\log s_{i,0}^t + \mu w_{i,1}^t\end{aligned}$$

I test the validity of this assumption using the sample of control users in Appendix C.

Proposition 5. *Assume,*

(A1) User behavior is in equilibrium at baseline, $s_{i,0}^t = p_i^t$

(A2) The system is in steady state, $\log s_{i,0}^t = \log s_{i,1}^t$

Let D_i *indicate treatment status. Then, for some updating parameter* θ*,*

$$\mathrm{E}\left[\log\left(\frac{s_{i,1}^t}{s_{i,0}^t}\right)\middle| D_i = 1\right] - \mathrm{E}\left[\log\left(\frac{s_{i,1}^t}{s_{i,0}^t}\right)\middle| D_i = 0\right] = \kappa - \theta \mathrm{E}[\log v_{i,0}^t | D_i = 1]$$

where, κ *is constant.*

Therefore, θ is identified using the relationship between the differences in toxic shares (from baseline $\tau = 0$ to intervention period $\tau = 1$) and the level of toxic views at baseline ($\tau = 0$) in the treated sample. The differences in sharing behavior between treatment and control groups account for unobserved heterogeneity.

6 Estimates

I use linear approximations of the log specification of the structural equation due to the presence of a large number of zeroes in the data on toxic views and shares (Chen and Yang, 2019). Figure D.19a shows that the structural relationship in Proposition 5 is approximately linear, and the estimated slope is negative.

6.1 OLS Estimates

Table E.4 shows that a 1% decrease in the proportion of toxic posts viewed during the intervention period decreases the proportion of toxic posts shared by $\hat{\theta}\% = 0.1\%$ only. This demonstrates stickiness in user behavior, as the elasticity in sharing behavior with respect to baseline exposure ($1 - \hat{\theta}$) is 0.9. These estimates support the claim that user behavior is not malleable, and is largely determined by user preferences at baseline.

However, the OLS estimates of θ are likely to be attenuated due to measurement error. This is because the treated users sample the toxic posts they view from a Binomial distribution because the full list of assigned posts is generated randomly, as each post can be toxic or not. I correct this using an IV strategy outlined in Appendix C.1.

6.2 IV Estimates

I instrument exposure to toxic content in the first half of the posts viewed at baseline, with the average toxicity in the second half to correct for measurement error. The setup requires an exclusion restriction which is satisfied because the sampling error in toxic posts viewed in the first half of posts viewed is uncorrelated to the sampling error in the second half of the posts viewed by construction. Then, the IV estimates in Column (2) of Table 3 indicate that the measurement error indeed attenuated the OLS estimates.

Column (1) in Table 3 shows the strength of the first stage in the IV specification. The corrected estimate in Column (2) shows that a 1% reduction in exposure to toxic content reduces engagement with toxic content by 0.16%. Therefore, the IV estimates indicate that the elasticity of sharing toxic content with respect to exposure at baseline is close to 0.84, and that user behavior significantly depends on pre-existing behaviors or preferences. I perform validation checks in Appendix C.2.

6.3 Model Based Counterfactuals

I calibrate the model parameters by matching moments of the empirical distribution of the total number of posts viewed and shared, as well as the total number of toxic posts shared.[26]

6.3.1 Alternative Behavioral Assumptions

I simulate the effects of diversification targeting more toxic users under different assumptions on user behavior. I describe the treatment effects, when users share the toxic content appearing on their feed mechanically ($\theta = 0$), and when users fully update their behavior in line with new information they are exposed to ($\theta = 1$).

Figure D.20 shows that the percentage change in number of toxic posts shared is decreasing in user preferences for toxic content when users are completely malleable. This is because more toxic users are more likely to be influenced when they view fewer toxic posts and when $\theta = 1$.

The decrease in number of toxic posts shared is larger in absolute terms when users are

[26] See Appendix C.3 for details.

fully malleable, than when the users behave according to the observed degree of malleability, $\theta = 0.16$. Finally, the model predicts that the number of toxic posts shared or the constituent parts of this measure, will not change for mechanical users with $\theta = 0$.

6.3.2 Model Based Decomposition

I decompose the treatment effect on the total number of toxic posts shared into two channels: (1) Engagement: change in platform usage, or the number of posts of any kind that were viewed and shared by treated users, and (2) Behavior: change in the probability of sharing toxic content for given exposure to toxic content. Algebraic manipulation of the main outcome variable gives

$$S_{i,1}^t = N_{i,1} \cdot \frac{S_{i,1}}{N_{i,1}} \cdot s_{i,1}^t$$

$$\implies \% \text{ change in } S_{i,1}^t = \underbrace{\% \text{ change in } N_{i,1}^t + \% \text{ change in } \frac{S_{i,1}}{N_{i,1}}}_{\text{change in engagement}} + \underbrace{\% \text{ change in } s_{i,1}^t}_{\text{change in behavior}}$$

Figure D.21a shows that the disengagement effect, or the decrease in overall platform usage, drives the decrease in the number of toxic posts shared when $\hat{\theta} = 0.16$. However, when users are fully malleable with $\theta = 1$, as in Figure D.21c, the treatment effect on number of toxic posts shared is driven by the change in proportion of shares that are toxic, s^t, or the behavioral changes due to the influence of exposure to diverse content. I also find that the (dis)engagement effect contributes to 55-60% of the total treatment effect, which is in line with the estimates from the empirical decomposition in Section 3.5.

6.3.3 Counterfactual Policies

Social media platforms frequently diversify user feeds by randomizing a portion of the posts that users see. This is because platforms typically want to be at some point on the exploration-exploitation frontier where they are able to retain users by showing them content they like, while continuously learning their potentially inconsistent preferences (Kleinberg et al., 2022). This paper shows that introducing diversity into feeds may also be beneficial from a societal viewpoint as it may persuade users to share less toxic content.

I simulate the main policy-relavant outcome, number of toxic posts shared, and its component parts in the model-based decomposition, under different mixes of algorithmic and random feeds in Figure 8. This shows that even when 60% of the feed is randomized, the effect on toxic sharing for toxic users is driven by behavior changes represented by s^t.

On the other hand, if at least 80% of the feed is randomized, the effect on toxic sharing

for toxic users is large and is driven by the engagement effect. This shows that a planner can optimally choose the degree of user feeds to diversify to balance the trade-off between user engagement with social media platforms, and the dissemination of toxic content.

7 Conclusion

In February 2019, a TikTok video of a drunk 28-year old man threatening to "butcher" villagers from an oppressed caste group surfaced on TikTok (Christopher, 2019). P Saravanan, the deputy inspector at Tiruttani police station in Tamil Nadu which dealt with this case said, "If you leave a gun on a table, it is partially your (TikTok) responsibility... What we have now is leaving a gun chest open." However, P Madhava Soma Sundaram, Professor of Criminology and Fellow of the Indian Society of Criminology pins more accountability on the demand for harmful content, "I will not blame Tiktok. This is a reflection of our society."

This paper studies the role of user preferences and personalization algorithms in driving engagement with extreme content. Using an individually randomized intervention with 8 million social media users in India, I examine whether the content presented by algorithms substantially impacts user choices, or if conversely, users seek out content consistent with their existing behavioral patterns. I show that the while the intervention significantly reduced user exposure to toxic content, there was an increase in the probability of sharing toxic posts conditional on viewing such content. A behavioral model rationalizes these results, and estimates show that the algorithm's influence on user behavior is limited. This leaves little room for policy instruments to alter sharing behavior through reduced exposure to toxic content.

I examine the mechanisms using observable user attributes and baseline behavior. Figure D.16 shows that, irrespective of treatment status, users with the highest affinity to toxic content at baseline (Q5) were more likely to **(1)** have adopted the platform early, **(2)** be less active at baseline but more active during working hours, **(3)** be male and older, **(4)** be more engaged with "politics," but less engaged with "greetings" at baseline.

Recall that, Q5 users disengage with the platform upon being treated, whereas Q1 users seek out content on the platform. I conjecture that this is because Q1 users log on to SM to consume content that is not found on other platforms (good morning messages, or other "greetings"). On the contrary, Q5 users are more likely to consume political content at baseline, which is substitutable on other platforms.

This is likely because the platform offers a unique opportunity for users to share posts directly to WhatsApp, making the platform a one-stop shop for posts that users want to

share. Moreover, there are no competing platforms that offer this type of 'WhatsApp-able' content in India as most content generation platforms (like Facebook, Instagram, or YouTube) encourage users to stay on their respective apps.[27] In Figure D.15, users with the highest affinity to greetings at baseline did not disengage with the platform. This may also explain the inelasticity of Q1 users.

Figure 5 presents survey evidence to show that treated users with higher affinity to toxic content spent more time on other platforms. On the other hand, no such trend was observed for users with lower affinity to toxic content. This shows that users with higher proclivity to toxic content at baseline are more likely to find such content on other platforms, when the intervention reduced their exposure on SM. Therefore, cross-platform regulation is necessary to effectively reduce the spread of toxic content.

This paper has the following limitations. First, it considers the effects of a specific algorithm on a specific platform. These results are generalizable to other platforms to the extent that they use similar algorithms to personalize content recommendations. Second, this analysis is restricted to the effects of the "random algorithm" for one month only. Future work will focus on understanding the long-term effects of the intervention, using administrative data for later months, and survey data for a random subset of users in the experimental sample. Finally, the broader implications of this intervention on mental health and digital addiction are also important to study, but were outside the scope of this paper. I aim to contribute to these strands of knowledge in future research work using the survey data on mental health outcomes that I collected at the end of the 11-month intervention period.

References

D. Acemoglu. Harms of ai. Technical report, National Bureau of Economic Research, 2021.

D. Acemoglu, D. Autor, J. Hazell, and P. Restrepo. Artificial intelligence and jobs: Evidence from online vacancies. *Journal of Labor Economics*, 40(S1):S293–S340, 2022.

D. Acemoglu, A. Ozdaglar, and J. Siderius. A model of online misinformation. *Review of Economic Studies*, page rdad111, 2023.

C. C. Aggarwal et al. *Recommender systems*, volume 1. Springer, 2016.

G. A. Akerlof and R. E. Kranton. Economics and identity. *The Quarterly Journal of Economics*, 115(3):715–753, 2000.

[27] In fact, this is the primary objective of the algorithm on these other platforms: to increase the time a user spends on the platform. See this Marketing guide `https://www.socialpilot.co/youtube-marketing/youtube-algorithm`

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

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

H. Allcott, M. Gentzkow, and L. Song. Digital addiction. *American Economic Review*, 112 (7):2424–63, 2022.

G. Aridor, D. Gonçalves, D. Kluver, R. Kong, and J. Konstan. The economics of recommender systems: Evidence from a field experiment on movielens. *arXiv preprint arXiv:2211.14219*, 2022.

G. Aridor, R. Jiménez-Durán, R. Levy, and L. Song. The economics of social media. *Journal of Economic Literatire*, 2024.

C. Arun. On whatsapp, rumours, lynchings, and the indian government. *Economic & Political Weekly*, 54(6), 2019.

E. Ash and S. Hansen. Text algorithms in economics. *Annual Review of Economics*, 15(1): 659–688, 2023.

S. Athey and G. W. Imbens. Machine learning methods that economists should know about. *Annual Review of Economics*, 11:685–725, 2019.

S. Athey, M. Bayati, N. Doudchenko, G. Imbens, and K. Khosravi. Matrix completion methods for causal panel data models. *Journal of the American Statistical Association*, 116(536):1716–1730, 2021.

M. Avalle, N. Di Marco, G. Etta, E. Sangiorgio, S. Alipour, A. Bonetti, L. Alvisi, A. Scala, A. Baronchelli, M. Cinelli, et al. Persistent interaction patterns across social media platforms and over time. *Nature*, 628(8008):582–589, 2024.

C. Bagchi, F. Menczer, J. Lundquist, M. Tarafdar, A. Paik, and P. A. Grabowicz. Social media algorithms can curb misinformation, but do they? *arXiv preprint arXiv:2409.18393*, 2024.

S. Banaji, R. Bhat, A. Agarwal, N. Passanha, and M. Sadhana Pravin. Whatsapp vigilantes: An exploration of citizen reception and circulation of whatsapp misinformation linked to mob violence in india. 2019.

J. Banerjee, J. N. Taroni, R. J. Allaway, D. V. Prasad, J. Guinney, and C. Greene. Machine learning in rare disease. *Nature Methods*, 20(6):803–814, 2023.

P. Barberá, J. T. Jost, J. Nagler, J. A. Tucker, and R. Bonneau. Tweeting from left to right: Is online political communication more than an echo chamber? *Psychological science*, 26 (10):1531–1542, 2015.

G. S. Becker. A note on restaurant pricing and other examples of social influences on price. *Journal of political economy*, 99(5):1109–1116, 1991.

G. Beknazar-Yuzbashev, R. Jiménez Durán, J. McCrosky, and M. Stalinski. Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*, 2022.

D. Björkegren, J. E. Blumenstock, and S. Knight. Manipulation-proof machine learning. *arXiv preprint arXiv:2004.03865*, 2020.

R. A. Blair, J. Gottlieb, B. Nyhan, L. Paler, P. Argote, and C. J. Stainfield. Interventions to counter misinformation: Lessons from the global north and applications to the global south. *Current Opinion in Psychology*, 55:101732, 2024.

L. Braghieri, R. Levy, and A. Makarin. Social media and mental health. *American Economic Review*, 112(11):3660–3693, 2022.

E. Brynjolfsson, A. Collis, A. Liaqat, D. Kutzman, H. Garro, D. Deisenroth, and N. Wernerfelt. The consumer welfare effects of online ads: Evidence from a 9-year experiment. *Available at SSRN 4877025*, 2024.

L. Bursztyn, D. Cantoni, D. Y. Yang, N. Yuchtman, and Y. J. Zhang. Persistent political engagement: Social interactions and the dynamics of protest movements. *American Economic Review: Insights*, 3(2):233–250, 2021.

L. Bursztyn, B. R. Handel, R. Jimenez, and C. Roth. When product markets become collective traps: The case of social media. Technical report, National Bureau of Economic Research, 2023.

L. Butera, R. Metcalfe, W. Morrison, and D. Taubinsky. Measuring the welfare effects of shame and pride. *American Economic Review*, 112(1):122–168, 2022.

Y. Chen and D. Y. Yang. The impact of media censorship: 1984 or brave new world? *American Economic Review*, 109(6):2294–2332, 2019.

C.-F. Chiang and B. Knight. Media bias and influence: Evidence from newspaper endorsements. *The Review of economic studies*, 78(3):795–820, 2011.

N. Christopher. Tiktok is fuelling india's deadly hate speech epidemic. Technical report, The Wired, 2019. URL `https://www.wired.com/story/tiktok-india-hate-speech-caste/`.

S. Coate and G. C. Loury. Will affirmative-action policies eliminate negative stereotypes? *The American Economic Review*, pages 1220–1240, 1993.

S. Dash, A. Arya, S. Kaur, and J. Pal. Narrative building in propaganda networks on indian twitter. In *Proceedings of the 14th ACM Web Science Conference 2022*, pages 239–244, 2022.

M. Dell. Deep learning for economists. *arXiv preprint arXiv:2407.15339*, 2024.

S. DellaVigna and E. Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

S. DellaVigna, J. A. List, and U. Malmendier. Testing for altruism and social pressure in charitable giving. *The Quarterly Journal of Economics*, 127(1):1–56, 2012.

E. Duflo, P. Dupas, and M. Kremer. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in kenya. *American economic review*, 101(5):1739–1774, 2011.

H. Fang and G. C. Loury. "dysfunctional identities" can be rational. *American Economic Review*, 95(2):104–111, 2005.

P. Fortuna, J. Soler, and L. Wanner. Toxic, hateful, offensive or abusive? what are we really classifying? an empirical analysis of hate speech datasets. In *Proceedings of the 12th language resources and evaluation conference*, pages 6786–6794, 2020.

F. Gao and L. Han. Implementing the nelder-mead simplex algorithm with adaptive parameters. *Computational Optimization and Applications*, 51(1):259–277, 2012.

M. Gentzkow and J. M. Shapiro. What drives media slant? evidence from us daily newspapers. *Econometrica*, 78(1):35–71, 2010.

M. Gentzkow and J. M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

A. Goldfarb and C. Tucker. Digital economics. *Journal of economic literature*, 57(1):3–43, 2019.

R. Gonzalez and E. M. Maffioli. Is the phone mightier than the virus? cellphone access and epidemic containment efforts. *Journal of Development Economics*, 167:103228, 2024.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. Reshares on social media amplify political news but do not detectably affect beliefs or opinions. *Science*, 381(6656):404–408, 2023a.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023b.

GWI. The global media landscape in 2023. Technical report, 2023.

GWI. Social media trends: 2024 global report. Technical report, 2024.

F. Haugen. Statement of frances haugen. *Sub-Committee on Consumer Protection, Product Safety, and Data Security*, 2021.

H. Hosseinmardi, A. Ghasemian, M. Rivera-Lanas, M. Horta Ribeiro, R. West, and D. J. Watts. Causally estimating the effect of youtube's recommender system using counterfactual bots. *Proceedings of the National Academy of Sciences*, 121(8):e2313377121, 2024.

F. Huszár, S. I. Ktena, C. O'Brien, L. Belli, A. Schlaikjer, and M. Hardt. Algorithmic amplification of politics on twitter. *Proceedings of the National Academy of Sciences*, 119 (1):e2025334119, 2022.

IIPS and ICF. National family health survey (nfhs-5), 2019-21, 2021.

ITA. India country commercial guide. *International Trade Administration*, 2024. URL `https://www.trade.gov/country-commercial-guides/india-information-and-communication-technology#:~:text=Overview,20%20percent%20of%20predicted%20GDP.`

C. Jaffrelot. Modi's india: Hindu nationalism and the rise of ethnic democracy. 2021.

R. Jiménez Durán. The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN*, 2022.

R. Jiménez Durán, K. Müller, and C. Schwarz. The effect of content moderation on online and offline hate: Evidence from germany's netzdg. *Available at SSRN 4230296*, 2023.

R. Jiménez Durán, K. Müller, and C. Schwarz. The effect of content moderation on online and offline hate: Evidence from germany's netzdg. *Available at SSRN 4230296*, 2024.

H. Jo and E.-M. Baek. Predictors of social networking service addiction. *Scientific Reports*, 13(1):16705, 2023.

A. Kalra. A'ghetto'of one's own: Communal violence, residential segregation and group education outcomes in india. 2021.

A. Kalra. Algorithmic drivers of behavior on social media. Technical report, AEA RCT Registry, 2023.

M. Kasy. Algorithmic bias and racial inequality: a critical review. *Oxford Review of Economic Policy*, 40(3):530–546, 2024.

M. Kasy. *How AI works, and for whom: A guide towards reclaiming democratic control*. University of Chicago Press, forthcoming.

J. Kleinberg, S. Mullainathan, and M. Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *arXiv preprint arXiv:2202.11776*, 2022.

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

D. S. Lee. Training, wages, and sample selection: Estimating sharp bounds on treatment effects. *Review of Economic Studies*, 76(3):1071–1102, 2009.

M. Manacorda and A. Tesei. Liberation technology: Mobile phones and political mobilization in africa. *Econometrica*, 88(2):533–567, 2020.

G. J. Martin and A. Yurukoglu. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9):2565–2599, 2017.

P. K. Masur, D. DiFranzo, and N. N. Bazarova. Behavioral contagion on social media: Effects of social norms, design interventions, and critical media literacy on self-disclosure. *Plos one*, 16(7):e0254670, 2021.

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

K. Müller and C. Schwarz. From hashtag to hate crime: Twitter and antiminority sentiment. *American Economic Journal: Applied Economics*, 15(3):270–312, 2023.

A. Narayanan and S. Kapoor. *AI snake oil: What artificial intelligence can do, what it can't, and how to tell the difference*. Princeton University Press, 2024.

B. Nyhan, J. Settle, E. Thorson, M. Wojcieszak, P. Barberá, A. Y. Chen, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, et al. Like-minded sources on facebook are prevalent but not polarizing. *Nature*, 620(7972):137–144, 2023.

Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019.

E. Pariser. *The filter bubble: What the Internet is hiding from you*. penguin UK, 2011.

S. M. Schennach. Recent advances in the measurement error literature. *Annual Review of Economics*, 8:341–377, 2016.

U. K. Schmid, A. S. Kümpel, and D. Rieger. Social media users' motives for (not) engaging with hate speech: An explorative investigation. *Social Media+ Society*, 10(4): 20563051241306322, 2024.

C. Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton university press, 2018.

A. K. Thakur. New media and the dalit counter-public sphere. *Television & New Media*, 21 (4):360–375, 2020.

A. Vaswani. Attention is all you need. *NeurIPS*, 2017.

A. Waghmare. Access to phones and the internet. 2024.

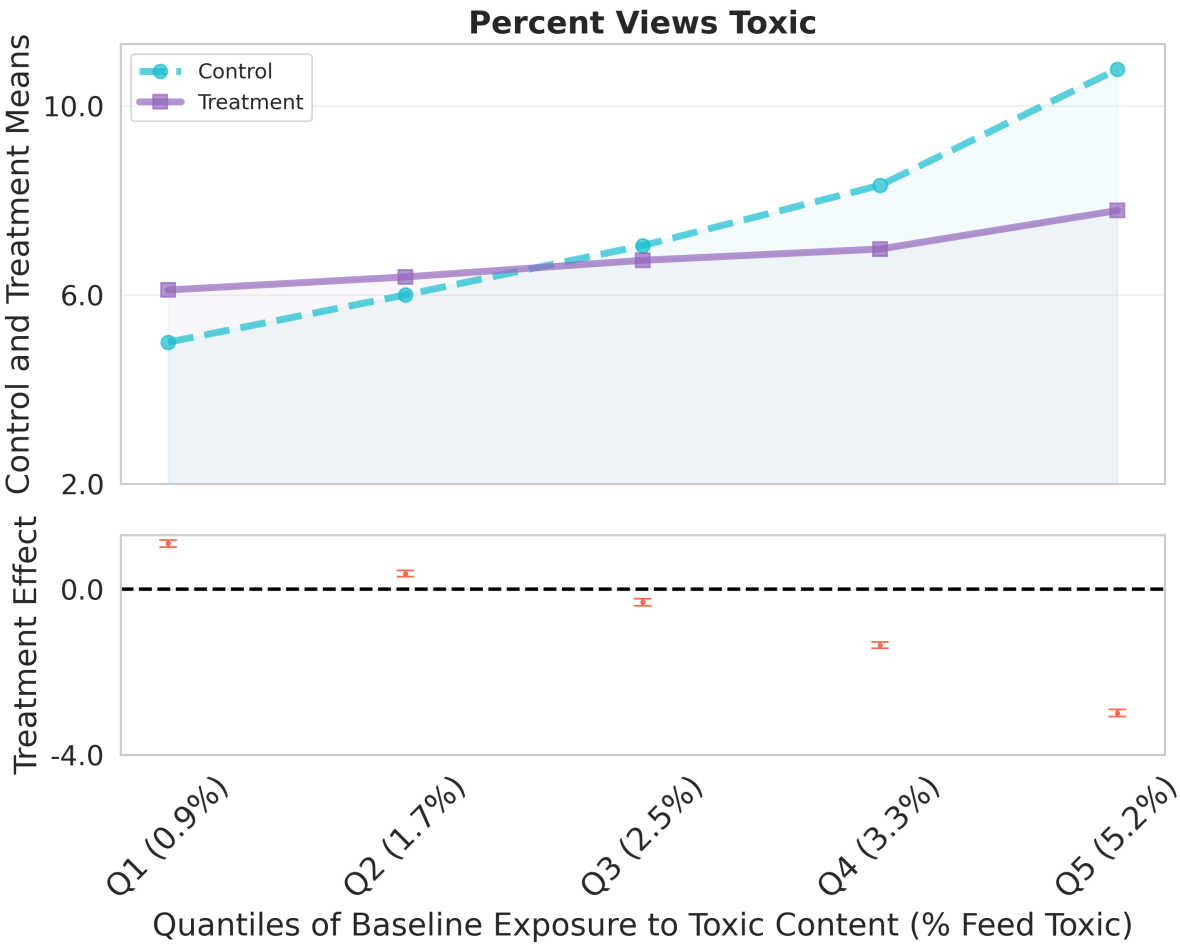
J. Waldron. Dignity and defamation: The visibility of hate. *Harv. L. Rev.*, 123:1596, 2009.

World Bank. Internet users as a share of the population, 2022. data retrieved from World Development Indicators, `https://databank.worldbank.org/source/world-development-indicators` on September 27, 2024.

E. Zhuravskaya, M. Petrova, and R. Enikolopov. Political effects of the internet and social media. *Annual review of economics*, 12:415–438, 2020.

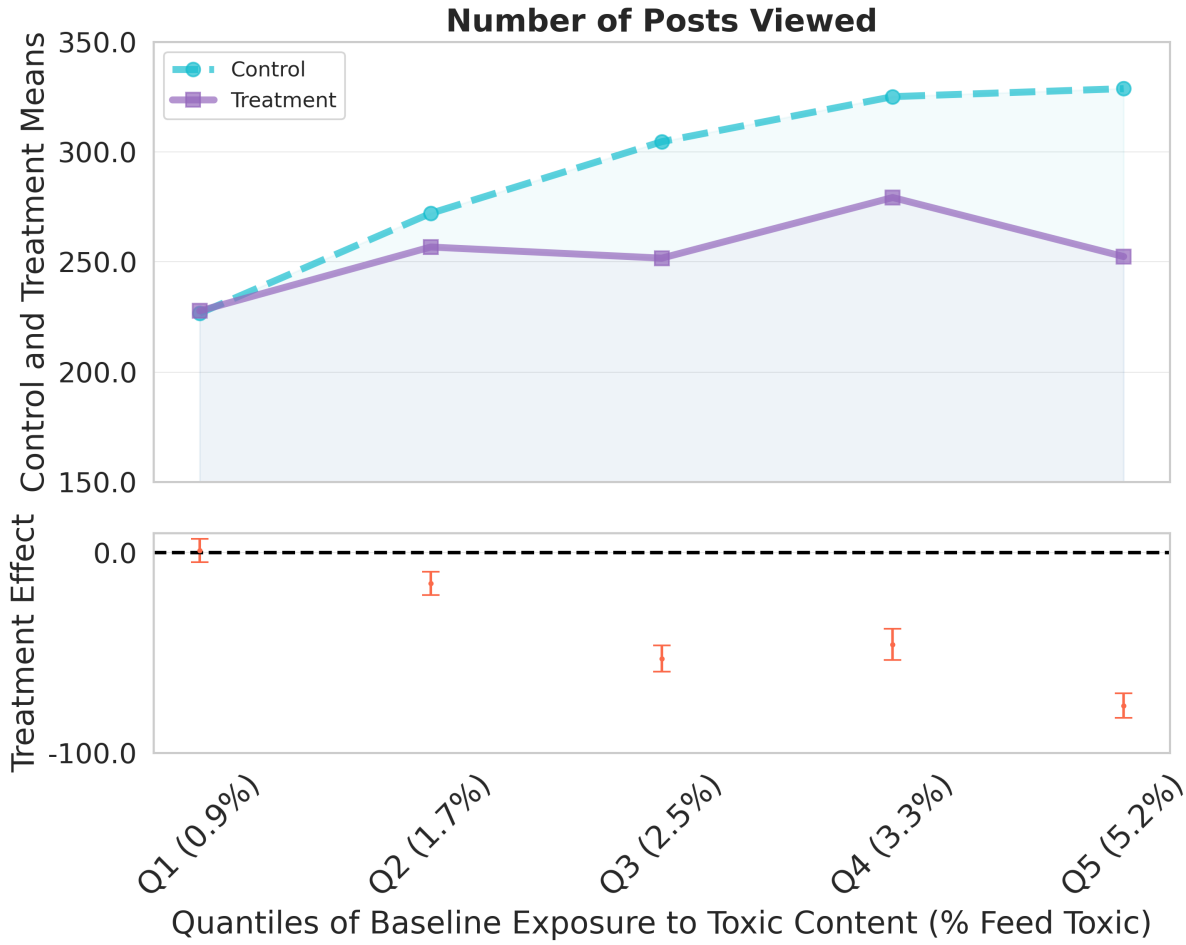
Tables and Figures

Figure 1: Treatment intensity by user type



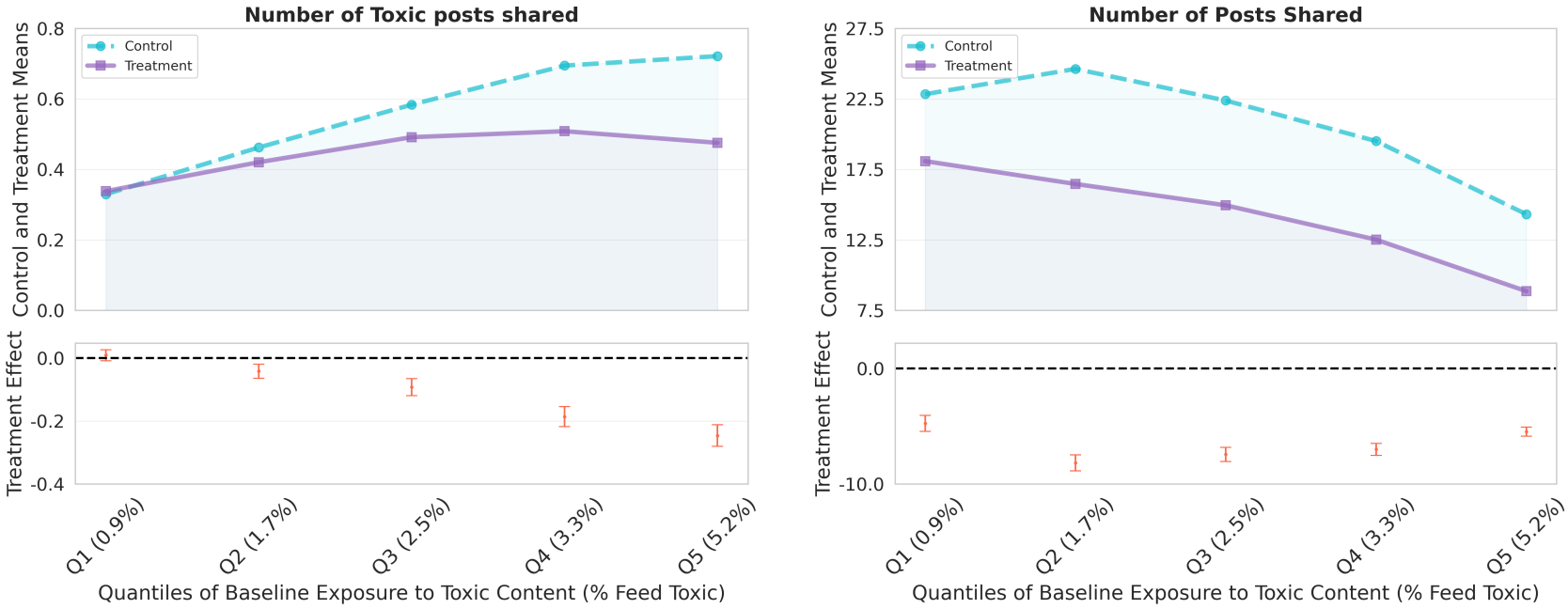
Notes: This figure shows that the treatment intensity, treatment effect on the percent toxic posts viewed, is decreasing with exposure to toxic posts at baseline. This effect is positive for Q1 users and negative for Q5 users, where each quantile is defined using the proportion of toxic posts viewed at baseline. The axis corresponding to the bottom plot shows the magnitude of the treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcome, percentage of posts viewed that are toxic, for each quantile. All regressions are run at the user level with robust standard errors.

Figure 2: Treatment effects on viewing behavior, by user type



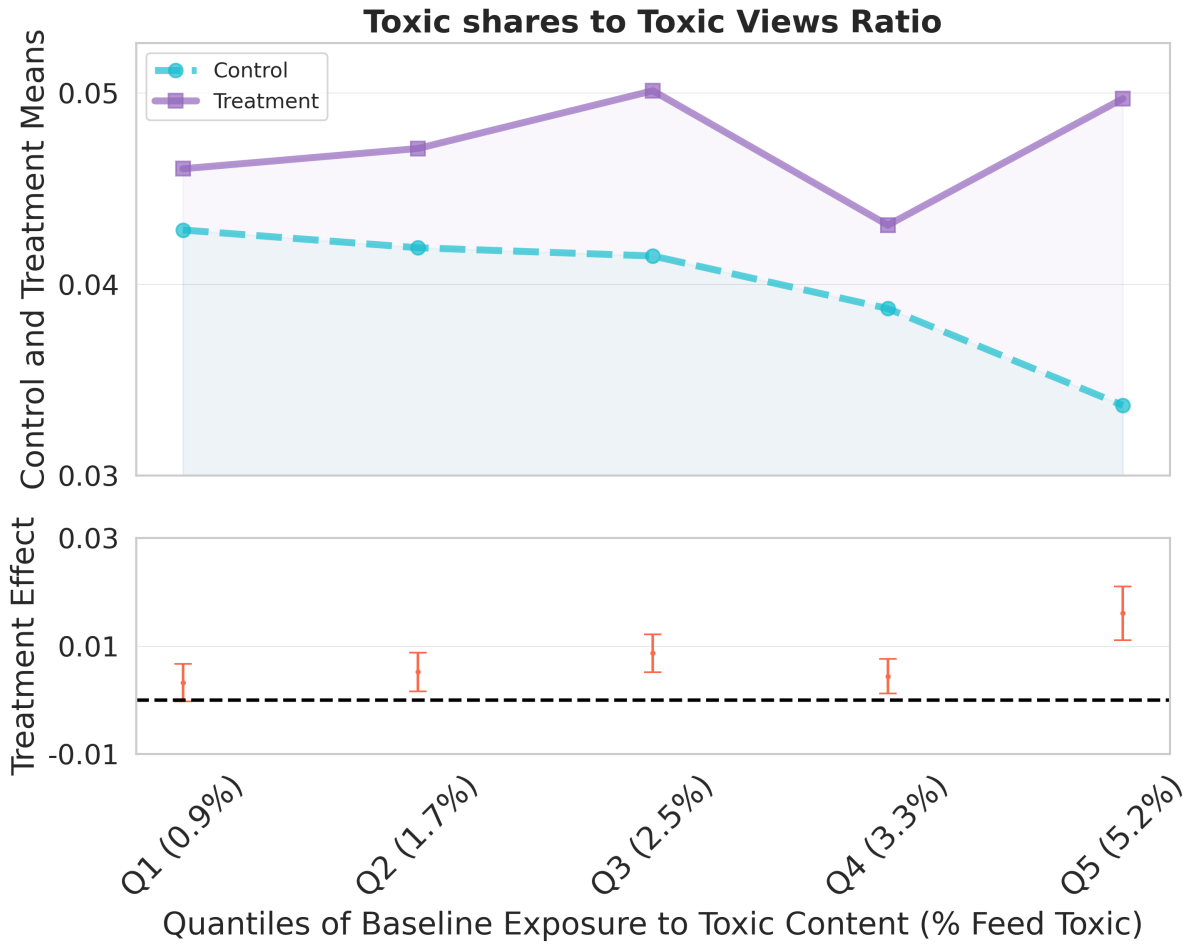
Notes: This figure shows that the total number of posts viewed, or overall engagement with the platform, also changes by treatment status and user type. In fact, the treatment effect on the total number of posts viewed is larger (in absolute terms) for users with higher exposure to toxic content at baseline. The axis corresponding to the bottom plots shows the magnitude of the treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions are run at the user level with robust standard errors.

Figure 3: Treatment effects on sharing behavior, by user type



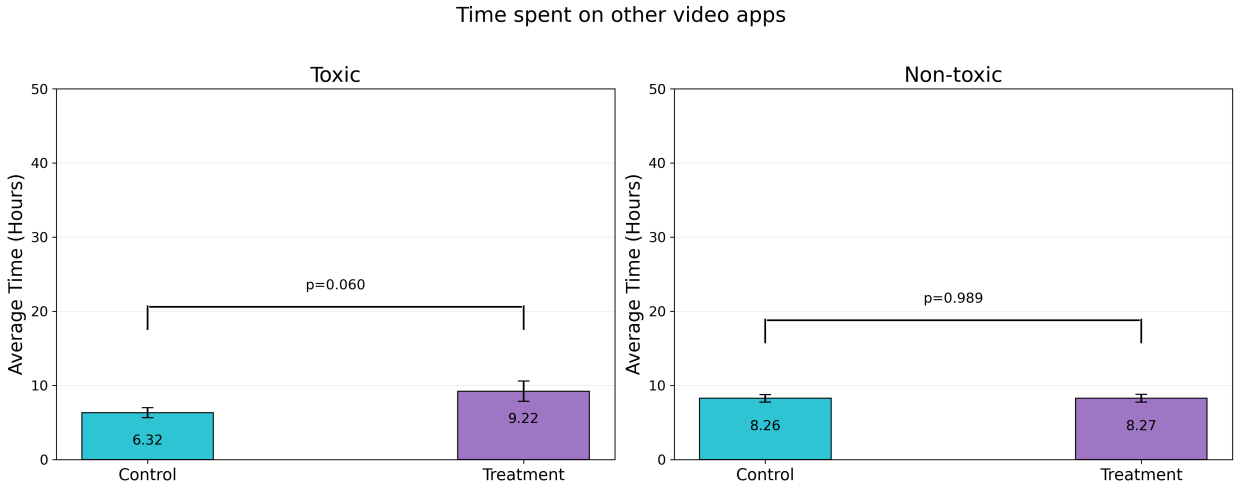
Notes: This figure shows that the treatment effect on the number of toxic posts shared is negative for toxic users (Q3–Q5) but is not statistically significant for Q1 users. The right panel shows that while the number of toxic posts shared by toxic users has decreased, these users disengage from the platform by sharing fewer posts overall. From this figure, it remains unclear whether toxic users share fewer toxic posts because they are exposed to less toxic content they can share, are influenced by the non-toxic content they are exposed to, or because they are disengaging with the platform. The axis corresponding to the bottom plots shows the magnitude of the treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions are run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure 4: Evidence on inelasticity in toxic sharing and seeking out behavior, by user type



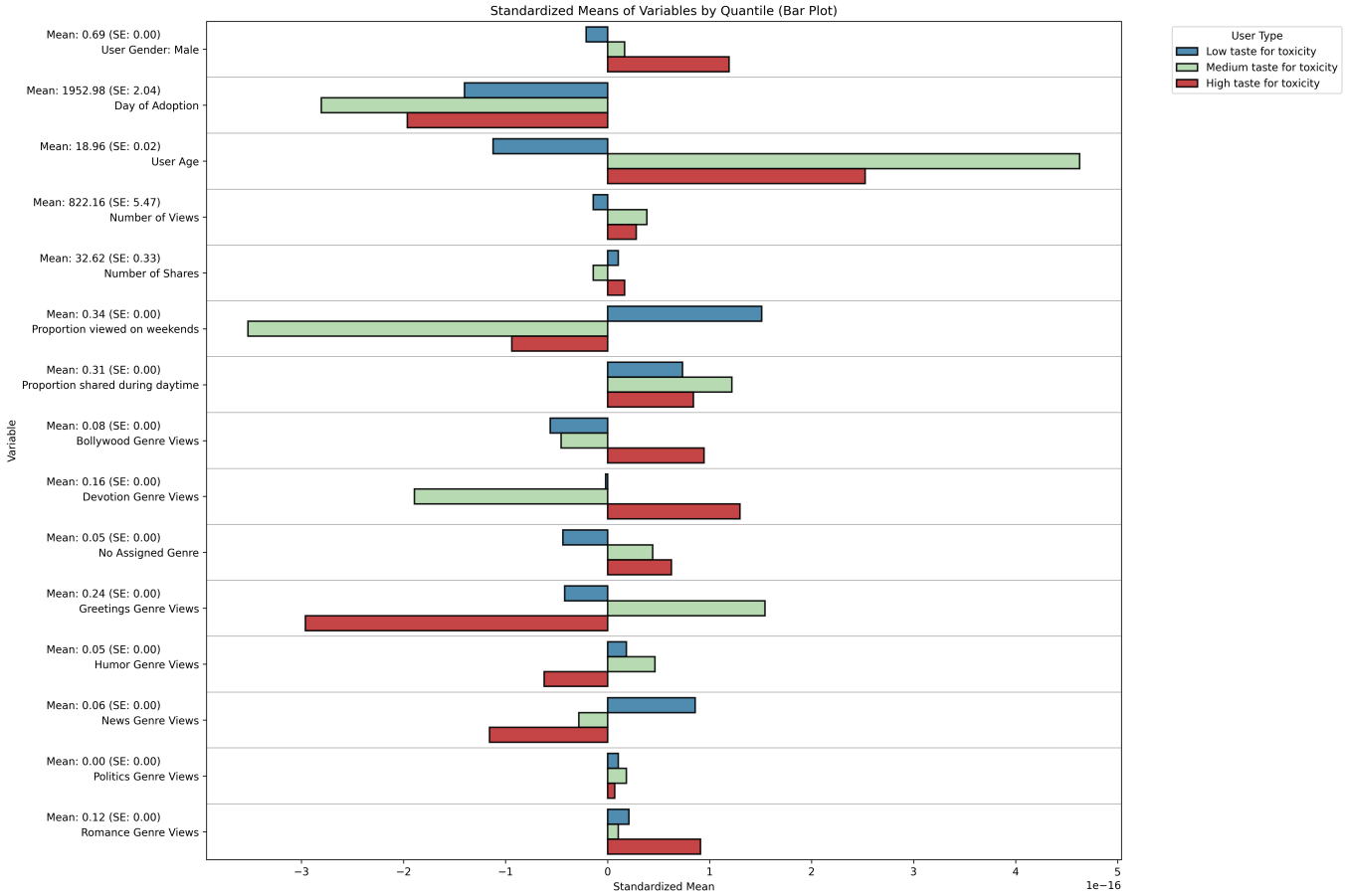
Notes: This figure shows that user behavior immalleable or "sticky". Treated users shared a higher proportion of the toxic posts they viewed during the intervention period, and the treatment effect on this measure is the largest for Q5 users. This suggests that the societal benefit of the policy is blunted as the reduction in the total number of toxic posts shared would have been larger if users had not shared a higher proportion of the toxic posts they viewed. The axis corresponding to the bottom plots shows the magnitude of the treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions are run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure 5: Substitution with other platforms



Notes: This Figure shows that users with higher proclivity to toxic content at baseline were more likely to spend more time on other platforms upon being treated (with a p-value of 0.06). A subset of users in the experimental sample were randomly selected for a follow-up survey ($N = 8,387$), and asked how much time they spent on a range of other social media platforms, and the time they spent on the TV, or telephone conversations, or in-person interactions. The survey was conducted at the end of the intervention period, with $4,236$ users randomly sampled from the treatment group, and the remaining $4,151$ users sampled from the control group.

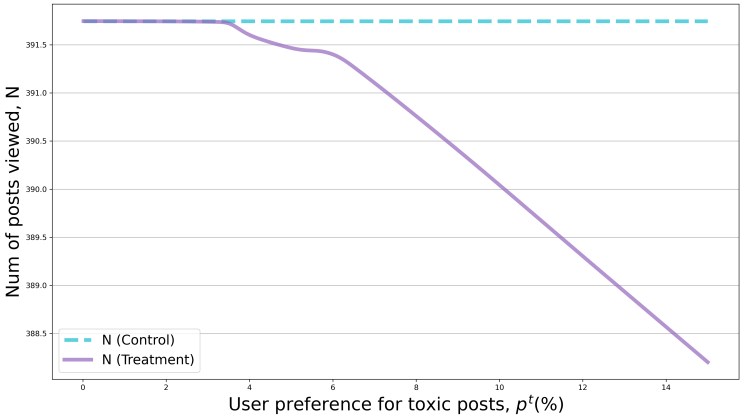
Figure 6: Key user attributes at baseline



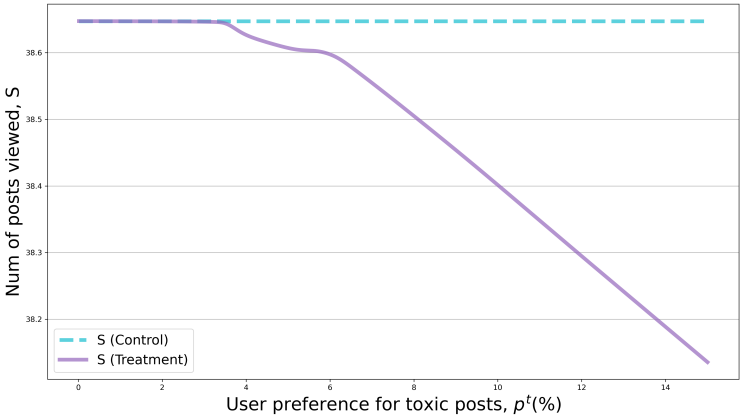
Notes: This Figure shows the baseline attributes of users in the experimental sample, distributed across user type. User type is defined by the proportion of toxic posts viewed at baseline, and users are allocated to the quantile in which they fall. The bar charts are constructed after standardizing the means of each variable. The means (and SEs) displayed with the name of each variable are not standardized.

Figure 7: Model predictions by user tastes for toxic content, p^t

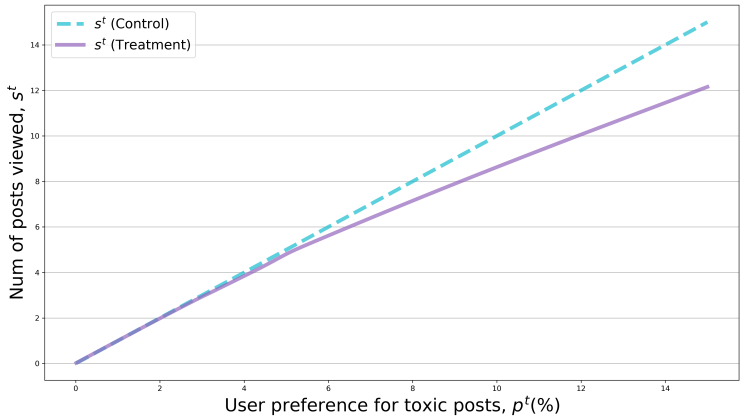
(a) Total number of posts viewed



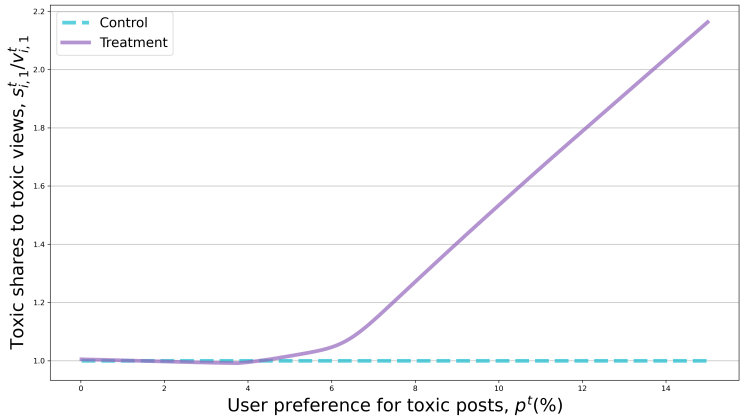
(b) Total number of posts shared



(c) Proportion of shares that are toxic

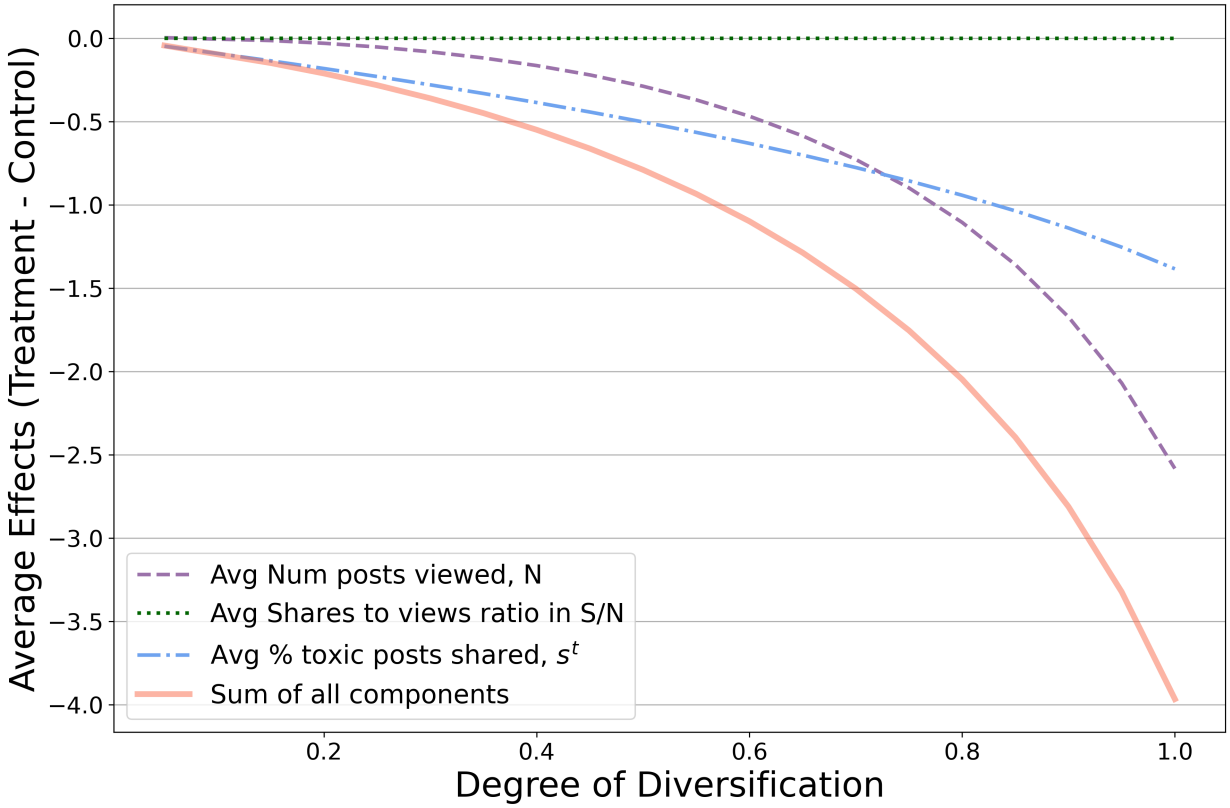


(d) Ratio of toxic shares to toxic views



Notes: This figure provides the model's predictions for key outcomes when the feed is randomized only for users with $p^t \geq \bar{q}^t$, where user type is defined by their tastes for toxic content, p^t. Panels (a) and (b) show that more toxic users (toward the right in the p^t distribution) are expected to view and share fewer posts upon being treated. Panel (c) shows that the treatment effect on the proportion of toxic shares is expected to be negative for toxic users. This is due to the larger reduction in total platform usage among toxic users and behavioral changes in the probability of sharing toxic content, both resulting from reduced exposure to such content. Panel (d) predicts that the ratio of toxic shares to toxic views is increasing in p^t. These predictions are obtained using calibrated parameters from the structural model by matching moment conditions for heterogeneous users.

Figure 8: Counterfactuals for different levels of randomization in content feeds



Notes: This figure simulates the counterfactual policy predictions for different levels of randomization in content feeds. The different degrees of randomization are achieved by considering linear combinations of the probabilities of being assigned toxic content in the control and treatment groups. That is, the counterfactual probabilities of being assigned toxic content under different policy regimes is given by $q_i^{t,a} = a \cdot \bar{q}^t + (1-a) \cdot q_i^t$. This shows that when $a = 60\%$, the decrease in the number of toxic posts is driven by the reduction in the probability of toxic users being assigned toxic content. This is ideal for a policymaker who wants to reduce the number of toxic posts viewed and shared without affecting the platform's overall engagement. However, as the degree of randomization increases to 80%, the reduction in toxic user engagement contributes more to the decrease in the number of toxic posts shared. Therefore, the policymaker can choose the degree of randomization, a, to balance this trade-off between reducing toxic engagement and maintaining overall engagement with the platform.

Table 1: Balance in treatment assignment across user characteristics and baseline behavior

Variable	Control Mean	Difference (T - C)	Std.Err.
Observable User Characteristics			
State: gujarat	0.021	-0.019	0.014
State: uttar pradesh	0.105	-0.012	0.012
City: aligarh	0.002	0.019	0.027
City: bareilly	0.002	-0.010	0.024
City: dehradun	0.001	0.012	0.028
City: faizabad	0.002	-0.038	0.026
City: hardoi	0.002	-0.020	0.025
City: jaunpur	0.003	-0.028	0.022
City: khandwa	0.001	-0.007	0.037
City: latur	0.001	-0.068	0.033
City: north east delhi	0.001	-0.054	0.034
City: pratapgarh	0.002	0.031	0.024
City: raipur	0.004	-0.005	0.023
City: sitapur	0.002	-0.017	0.026
Gender: Male	0.699	-0.002	0.003
Age: 19-30	0.006	0.000	0.016
Week: 2016-28	0.000	-0.662	10.698
Week: 2022-38	0.012	-0.748	10.696
Baseline Behavior			
Num Posts Viewed	777.126	0.000	0.000
Num Posts Shared	22.045	-0.000	0.000
Num Logins	9.250	-0.000	0.000
Time Spent (in hours)	16.341	-0.000	0.000
Prop Activity during Daytime	0.097	-0.001	0.004
Prop Activity during Weekends	0.346	-0.007	0.005
Num Searched per Post Viewed	0.175	0.001	0.002
Prop Views in Humor Genre	0.051	-0.009	0.030
Prop Views in News Genre	0.058	-0.008	0.030
Prop Shares in Bollywood Genre	0.010	-0.037	0.012
Prop Shares in News Genre	0.009	-0.010	0.014
Prop of Views Toxic (%)	2.681	0.007	0.007
Prop of Shares Toxic (%)	2.241	-0.029	0.042
Tox Share to Tox View Ratio	1.023	-0.000	0.000
F-statistic:	0.984	p-value:	0.506
N		231814	

Notes: This table shows balance in treatment assignment across all observable characteristics, using a single regression run at the user level: $D_i = \beta_0 + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where D_i is a binary variable equal to 1 when user i was assigned to the treatment group. The table shows coefficients corresponding to a randomly selected set of user characteristics, with weeks representing the date on which a user created their account. Additionally, none of the observable characteristics are correlated with treatment assignment. I cannot reject the null hypothesis of joint insignificance, with an F-statistic of 0.984 and a p-value of 0.506. The regression is estimated at the user level, and robust standard errors are in parentheses.

Table 2: Experimental Effects on Post Views and Shares

	Num Posts Viewed	Num Posts Shared
Treatment Effect	-35.497**	-6.367**
	(2.208)	(0.206)
Control Mean	246.654**	18.396**
	(1.361)	(0.131)
	Num Toxic Posts Viewed	Num Toxic Posts Shared
Treatment Effect	-5.024**	-0.093**
	(0.172)	(0.010)
Control Mean	18.806**	0.474**
	(0.129)	(0.006)
	% Toxic Posts Viewed	Toxic Shares to Views Ratio
Treatment Effect	-0.641**	0.007**
	(0.033)	(0.001)
Control Mean	7.416**	0.040**
	(0.018)	(0.001)
N	231814	

Notes: This table shows that the treatment led to overall disengagement with the platform, decrease in the total number of toxic posts viewed and shared, but an increase in the rate at which users shared toxic content they viewed. I find that the treatment effect on the number of posts viewed and shared, and the number of toxic posts viewed and shared, in one month is negative and statistically significant. This suggests that the treatment effect on engagement with toxic content would have been larger if users had not shared a higher proportion of the toxic posts they viewed. Each cell reports estimates of the regression coefficient from a linear regression of the outcome aggregated at the user level over all days in the first month of the intervention period (February 10 to March 10, 2023). Robust standard errors are in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table 3: Structural estimation of influence parameter θ, with measurement error correction

	(1)	(2)	(3)
	Proportion of Toxic Posts Viewed (Baseline, half-2)	Proportion of Toxic Posts Shared (Intervention - Baseline)	
Proportion of Toxic Posts Viewed (Baseline, half-1)	0.572***	-0.155**	
	(0.004)	(0.058)	
Proportion of Toxic Posts Viewed (Baseline, half-2)			-0.183**
			(0.065)
N	63041	63041	63041

Notes: This table provides estimates for the structural parameter θ in the model of sharing behavior, where θ captures the rate at which users update their behavior according to perceived social norms. θ is therefore the influence effect of one month's exposure to non-personalized feeds. Column (1) shows the relevance of the instrument, i.e., the proportion of toxic posts viewed computed using only the first half of posts viewed by a user at baseline, when they were arranged in a random order ($half1$). This instrument is used to correct the measurement error resulting from treated users randomly sampling toxic posts from their feeds. The independent variable in Column (1) is the proportion of toxic posts viewed at baseline, computed using only the second half of posts viewed by a user at baseline, when they were arranged in a random order ($half2$). Column (2) shows the results of an 2SLS regression on the first difference in proportion of toxic posts shared between baseline and the intervention period. Here, the independent variable is $half1$, which is instrumented with $half2$. Column (3) estimates the model with classical measurement error correction in STATA, where the correlation between $half1$ and $half2$ serves as the reliability measure for the proportion of toxic posts viewed. The estimated slope coefficient estimates γ_1 is always negative and statistically significant. Consequently, the estimated θ is positive, with a value of 0.16 based on the preferred specification in Column (2). The baseline period is December 2022, and the intervention period data span from February 10, to March 10, 2023. Robust standard errors are in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

A Details of Behavioral Model

This theoretical framework outlines users' incentives to view and share different types of content. That is, **(1)** users have some innate tastes for toxic content, and **(2)** they want to signal their type to conform with society's tastes (as perceived by the user). Users are assumed to update their perception of norms based on the content they view, and the algorithm in turn internalizes distortions from users' desire to conform to social norms.

A.1 Equilibrium

I solve for the subgame perfect equilibrium, and introduce user (i) and time (τ) subscripts. All four stages of the game are assumed to be played in sequence, in both the time periods, $\tau = 0$ (baseline) and $\tau = 1$ (intervention period). By backward induction, users first maximize utility by choosing the total number of posts to share, and also the number of toxic posts to share, i.e. $S_{i,\tau}$ and $S^t_{i,\tau}$, respectively. Users' best response characterizes one of the main outcome variables, i.e. proportion of toxic posts to share, $s^t_{i,\tau} = S^t_{i,\tau}/S_{i,\tau}$.

Lemma A.1. *For a utility maximizing agent i,*

$$s^t_{i,\tau} = \left(q^t_{i,\tau}\right)^{\theta}\left(p^t_i\right)^{1-\theta} \tag{8}$$

That is, users place a weight of θ social norms, as perceived by the user through her feed, while choosing the proportion of posts shared that are toxic.

Proof. The claims follow from users' first order condition (with respect to $s^r_{i,\tau}$) from the utility maximization problem. □

The optimal sharing strategy is a combination of user's own tastes and the content she is shown, $q^t_{i,\tau}$, weighted by $(1-\theta)$ and θ, respectively. The distribution of toxic posts on a user's content feed informs her about the type of content that a similar user is engaging with, and is therefore, socially acceptable. She values conformity with these perceived norms according to some factor θ. Otherwise, sharing decisions are made according to the user's own immutable tastes for toxic content, p^t_i. The user also decides the number of non-toxic posts she will share, if any, upon viewing posts in their feed.

Lemma A.2. *For a utility maximizing agent i,*

$$S_{i,\tau} = \frac{1}{2(\eta+\alpha)}\left[2N_{i,\tau}\alpha - \delta\theta(1-\theta)\left(\left(\log p^t_i\right)^2 - 2\log q^t_{i,\tau}\log p^t_i + \left(\log q^t_{i,\tau}\right)^2\right)\right] \tag{9}$$

That is, total number of posts shared is higher for more engaged users, with higher $N_{i,\tau}$; but is decreasing in the cost of sharing, η and the cost of viewing content that is not shareable, α.

Proof. The SPE's are solved for using backward induction. This follows from the first order condition of the user's utility maximization problem, after substituting optimal $s^t_{i,\tau}$ in the utility function. □

The number of posts shared seems to be increasing in user's own taste for toxic content, p_i^t, as well as their perception of society's tastes, conveyed by $q_{i,\tau}^t$. However, the correct comparative statics with respect to $S_{i,\tau}$ take into account the fact that total shares depend on the endogenous response to the total number of posts viewed N. Then, a forward-looking rational user i solves for the total number of posts to view, N, or the total time she spends on the platform looking at posts.

Lemma A.3. *For a utility maximizing agent i,*

$$N_{i,\tau} = \frac{1}{2\alpha\eta}\left[\beta(\alpha+\eta) - \delta\alpha\theta(1-\theta)\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2\right] \tag{10}$$

That is, users view a smaller number of posts when there is a mismatch between their preferences and the algorithmically generated preferences, $q_{i,\tau}^t \neq p_i^t$.

Proof. I begin by substituting the optimal sharing behavior (from Lemmas A.1 and A.2) into the utility function. User's first order condition, with respect to the total number of posts viewed generates the required expression. This shows that $N_{i,\tau}$ is decreasing in $\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2 > 0$. Therefore, $N_{i,\tau}$ is maximized when $q_{i,\tau}^t = p_i^t$. □

This clearly shows that when users are assigned content randomly, they are likely to spend less time on the platform. This is because the recommendations do not match user preferences, as extreme reated users are recommended the average user's feed. Lemma A.3 describes the total number of posts viewed in terms of the model's primitives. Subsequently, $N_{i,\tau}^t$ in equilibrium helps in determining the total number of posts shared, $S_{i,\tau}$. The given utility form provides two solutions for the total number of posts shared, one of which is zero. I describe the non-zero solution in terms of model primitives.

Lemma A.4. *For a utility maximizing agent i,*

$$S_{i,\tau} = \frac{1}{2\eta}\left[\beta - \delta\theta(1-\theta)\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2\right] \tag{11}$$

That is, users share a smaller number of posts when there is a mismatch between their preferences and the algorithmically generated preferences, $q_{i,\tau}^t \neq p_i^t$.

Proof. This expression is obtained by substituting (10) into the optimal sharing function in (11). This shows that $S_{i,\tau}$ is decreasing in $\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2 > 0$. Therefore, $S_{i,\tau}$ is maximized when $q_{i,\tau}^t = p_i^t$. □

The solution to the user's problem is therefore, fully characterized for the given probability of being assigned toxic content, $q_{i,\tau}^t$. For the given timing of the game, I finish characterizing the equilibrium by solving for the algorithm's optimal assignment probabilities. The platform's customization algorithm is trained to maximize the expected number

of posts viewed in order to increase eyeballs on advertisement posts that are interspersed on the users' ranked content feed. Therefore, the platform feeds the objective function in (10) to the algorithm, which in turn optimally chooses $q^t_{i,\tau}$ to maximize advertisement revenues.

Lemma A.5.

$$q^t_{i,\tau} = p^t_i \tag{12}$$

That is, the algorithm assigns toxic posts with probability equal to user's intrinsic tastes for toxic content.

Proof. This follows directly from the first order conditions of an algorithm that is set to maximize $N_{i,\tau}$ in (10), by choosing $q^t_{i,\tau}$ optimally. The same result follows if the algorithm's objective is defined more broadly, choosing $q^t_{i,\tau}$ to maximize $N^t_{i,\tau}$, or $S^t_{i,\tau}$, or some linear combination of the two. This is because the number of posts viewed and shared is decreasing in $\left(\log \frac{q^t_{i,\tau}}{p^t_i}\right)^2 \geq 0$, which equals zero when $q^t_{i,\tau} = p^t_i$. □

Recall that the assignment probabilities provide a heuristic for the algorithm, that provides an intuitive explanation for what the algorithm actually does. This intuitive result shows that the algorithm caters to users' intrinsic tastes for viewing toxic content. The algorithm internalizes users' incentives to signal their type and their conformity, but in equilibrium the algorithm assigns toxic content according to user's intrinsic tastes.[28] The result provides concrete basis to analyze behavior according to user type, where types are characterized according to the proportion of toxic posts assigned to them at baseline. This is because, in the equilibrium at baseline, the assignment probabilities are necessarily equal to user's intrinsic tastes for toxic content. The model provides comparative statics, that generate implications tested in the data.

These results demonstrate how different type of users respond differently to the treatment. If users were mechanical, they would all have the same behavioral response such that the proportion of toxic posts shared is equal to the proportion of toxic posts viewed, irrespective of treatment status and time period. The comparative statics show that the treatment effects on toxic sharing are unlikely to be mechanical. This means that users put a positive weight on the new information they receive when making sharing decisions. The model predicts mechanical behavior if and only if the influence parameter, θ equals 0 for mechanical users. I show that this parameter is non-trivial.

Lemma A.6. *User i with $N_{i,\tau}, S_{i,\tau} > 0$, is said to behave 'mechanically' when*

$$\theta = \beta = \eta = 0$$

That is, when $\theta = 0$, the elasticity of the proportion of toxic posts shared with the respect to the proportion of toxic posts viewed is 1.

[28] That is, the algorithm enables the self-fulfilling prophecy characteristic of statistical discrimination models, where user types determine the type of content users are assigned, and users share these posts to in turn, signal their type (Coate and Loury, 1993).

Proof. If, $\theta = 0$, the utility maximization problem becomes,

$$\max_{s^t,S,N} = \alpha(N-S)^2 - \delta S\left(\log\frac{s^t}{p^t}\right)^2 - \eta S^2 \tag{13}$$

Utility is maximized with respect to s^t when $s^t_{i,\tau} = p^t_i$. Then, by definition,

$$\frac{S^t_{i,\tau}}{S_{i,\tau}} = s^t_{i,\tau} = p^t_i$$

We know that in equilibrium, $q_{i,\tau} = p^t$. Then, assuming users view all the posts they are assigned, we have, $N^t_{i,\tau} = q^t_{i,\tau}N$. Therefore,

$$\frac{S^t_{i,\tau}}{S_{i,\tau}} = \frac{N^t_{i,\tau}}{N_{i,\tau}} = p^t_i = q^t_{i,\tau} \tag{14}$$

Then the treatment implies that,

$$\frac{\partial s^t_{i,\tau}}{\partial \bar{q}^t} = \frac{\partial v^t_{i,\tau}}{\partial \bar{q}^t} = 1 \tag{15}$$

where, $v^t_{i,\tau} = \frac{N^t_{i,\tau}}{N_{i,\tau}}$, and, elasticity of toxic sharing with respect to toxic viewing is

$$\frac{\partial s^t_{i,\tau}/\partial \bar{q}^t}{\partial v^t_{i,\tau}/\partial \bar{q}^t} = 1$$

□

When $\theta = 0$, users are considered mechanical as they share a fixed proportion of toxic content they view in each time period. The negation of this implication is also true, and is tested empirically to analyze if user behavior is malleable or sticky. That is, if users do not behave mechanically, then exposure has an influence on user behavior, i.e. $\theta > 0$.

I find that the treatment effect on the proportion of toxic posts shared is distinct from the effect on the proportion of toxic posts viewed. In stating Fact III before, I rejected the hypothesis that the elasticity of toxic sharing with respect to toxic viewing equals 1, with a p-value of 0.002. This shows that there are behavioral responses to diversifying content feeds, even though the influence of exposure is relatively small, as I show with the estimated model parameters.

The prediction helps negate the possibility that $\theta = 0$, which is the case of no updating in user behavior with respect to exposure. This sets up the rationale for estimating the structural model.

B Proofs for Theoretical Framework

Proof of Proposition 1

Let D_i be a binary variable indicating treatment status, so that $D_i = 1$ for treated users. Then, for users with $p_i^t > \bar{q}^t > 0$,

$$v_{i,1}^t(D_i = 1) - v_{i,1}^t(D_i = 0) = \bar{q}^t - q_{i,\tau}^t < 0$$

where, $v_{i,1}^t$ is the proportion of posts viewed that are toxic for user i at $\tau = 1$.

Proof. $\bar{q}^t - q_{i,\tau}^t < 0$, for users with higher baseline exposure to toxic content with $p_i^t > \bar{q}^t$ as $\bar{q}^t$ is the average user's probability of being assigned toxic content. Then, assuming users view everything they are assigned, $v_{i,1}^t(D_i = 1) = \bar{q}^t$. The fact that $v_{i,1}^t(D_i = 0) = q_{i,1}^t = p_i^t$ under the equilibrium condition for the control group completes the proof. □

Proof of Proposition 2

For user i with $\alpha, \eta, N_{i,\tau} > 0$, and $p_i^t > \bar{q}^t$,

$$\frac{\partial^2 N_{i,\tau}}{\partial p_i^t \partial \bar{q}^t} \geq 0$$

That is, the reduction in the total number of posts viewed, on account of the treatment, is larger for users with higher proclivity to toxic content.

Proof. Lemma A.3 implies

$$N_{i,\tau} = \frac{1}{2\alpha\eta}\left[\beta(\eta + \alpha) - \delta\alpha\theta(1-\theta)\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2\right]$$

With random content assignment during the intervention period ($\bar{q}^t$),

$$\frac{\partial N_{i,\tau}}{\partial \bar{q}^t} = \frac{-1}{2\alpha\eta}\left[\frac{2}{\bar{q}^t}\delta\alpha\theta(1-\theta)\log\frac{\bar{q}^t}{p_i^t}\right]$$

Note that, $p_i^t > \bar{q}^t$ is both necessary and sufficient for the derivative to be positive. That is, for users with higher proclivity to toxic content, randomly increasing the probability of assigning such content increases the number of posts viewed. Consider, the cross derivative with respect user tastes, p_i^t gives,

$$\frac{\partial^2 N_{i,\tau}}{\partial p_i^t \partial \bar{q}^t} = \frac{1}{2\alpha\eta}\left[\frac{2}{\bar{q}^t p_i^t}\delta\alpha\theta(1-\theta)\right] \geq 0$$

because $\theta \in [0,1]$, $\bar{q}^t, p_i^t \in (0,1)$, and $\alpha, \eta, \beta, \delta > 0$.

Then, for $p_i^t > \bar{q}^t$, random increases in probability of assigning toxic content increases the number of posts viewed, and the increase is larger for more toxic users. Conversely, when exogenous reductions in $\bar{q}^t$ decrease the number of posts viewed for toxic users, the reduction is larger for more toxic users. □

Proof of Proposition 3

For user i with $\eta, N_{i,\tau}, S_{i,\tau} > 0$,

$$\frac{\partial^2 s_{i,\tau}^t}{\partial p_i^t \partial \bar{q}^t} \geq 0$$

That is, the treatment effect on the proportion of toxic posts shared is negative and smaller for users with higher proclivity to toxic content.

Proof. From Lemma A.1 shows that,

$$s_{i,\tau}^t = \left(q_{i,\tau}^t\right)^{\theta} \left(p_i^t\right)^{1-\theta}$$

Then, we can see that

$$\frac{\partial^2 s_{i,\tau}^t}{\partial p_i^t \partial \bar{q}^t} = \theta(1-\theta)\left(q_{i,\tau}^t\right)^{\theta-1}\left(p_i^t\right)^{-\theta} \geq 0$$

for $\theta \in [0,1]$, and $q_{i,\tau}^t, p_i^t \in (0,1)$. □

Proof of Proposition 4

For user i with $\alpha, \eta > 0$, $\theta \in [0,1]$ and $p_i^t > \bar{q}^t > 0$,

$$\frac{\partial^2 (s_{i,\tau}^t / \bar{q}^t)}{\partial p_i^t \partial \bar{q}^t} \leq 0 \tag{16}$$

That is, marginal increases in the average probability of assigning toxic content leads to smaller changes in the proportion of toxic shares out of toxic views for users who prefer such content.

Proof. As before,

$$\frac{\partial^2 (s_{i,\tau}^t / \bar{q}^t)}{\partial p_i^t \partial \bar{q}^t} = -(1-\theta)^2 q^{\theta-2} p^{-\theta} \leq 0$$

□

Proof of Lemma B.1

Lemma B.1. *Estimates of θ from the relationship between sharing behavior and the proportion of toxic content viewed during the intervention period among a sample of control users is not identified.*

Proof. Consider the linear structural relationship,

$$\log s_{i,1}^t - \log s_{i,0}^t = \theta v_{i,1}^t + \log w_i^t$$

and suppose that, *by contradiction*, θ is estimable, using control users. This necessarily implies that $\mathrm{E}[\log w_i^t | \log v_{i,1}^t] = 0$. The steady state condition implies that the left-hand side of the equation is always zero, meaning

$$\mathrm{E}[\log s_{i,1}^t - \log s_{i,0}^t | \log v_{i,1}^t] = 0$$

This implies that $\theta = 0$. However, this contradicts Lemma A.6 which shows that $\theta > 0$. Therefore, θ is not estimable from this relationship, in the sample of control users. □

Proof of Proposition 5

Assume,

(A1) User behavior is in equilibrium at baseline, $s_{i,0}^t = p_i^t$

(A2) The system is in steady state, $\log s_{i,0}^t = \log s_{i,1}^t$

Then, for some updating parameter θ for all treated users i,

$$\mathrm{E}\left[\log\left(\frac{s_{i,1}^t}{s_{i,0}^t}\right) \middle| D_i = 1 \right] - \mathrm{E}\left[\log\left(\frac{s_{i,1}^t}{s_{i,0}^t}\right) \middle| D_i = 0 \right] = \kappa - \theta \mathrm{E}[\log v_{i,0}^t | D_i = 1]$$

where, $\bar{q^t}$ is constant, and D_i indicates treatment status.

Proof. Lemma A.1 gives the optimal sharing function,

$$s_{i,1}^t = \left(v_{i,1}^t(q_{i,1}^t)\right)^\theta \left(s_{i,0}^t\right)^{1-\theta} e^{\mu w_{i,1}^t} \tag{17}$$

where, $w_{i,1}^t$ is some *iid* taste-based shock. Then, the steady state condition gives

$$s_{i,0}^t = s_{i,1}^t\left(v_{i,1}^t(q_{i,1}^t), s_{i,0}^t, w_{i,1}^t\right) \tag{18}$$

$$\implies s_{i,0}^t = v_{i,1}^t e^{w_{i,1}^t \frac{\mu}{\theta}} \tag{19}$$

Substituting $s_{i,0}^t$ into (17) for control users,

$$s_{i,1}^t = v_{i,1}^t e^{w_{i,1}^t \frac{\mu}{\theta}} \text{ and } s_{i,0}^t = v_{i,0}^t e^{w_{i,0}^t \frac{\mu}{\theta}} \tag{20}$$

On the other hand, for treated users, substituting the steady state condition (18) into the optimal sharing function (17) gives

$$s_{i,1}^t = \left(v_{i,1}^t\right)^\theta \left(\bar{q^t}\right)^{1-\theta} e^{w_{i,1}^t \frac{\mu}{\theta}} \tag{21}$$

Moving subscripts to previous period, we get

$$s^t_{i,0} = \left(v^t_{i,0}\right)^{\theta}\left(\bar{q^t}\right)^{1-\theta} e^{w^t_{i,0}\frac{\mu}{\theta}} \tag{22}$$

Let, D_i indicate treatment status of user i. Then, plugging in values from equations (18), (21), and (22), into the difference in the difference in sharing behavior across $\tau = 1$ and $\tau = 0$ across treatment and control groups,

$$\mathrm{E}\left[\log\left(\frac{s^t_{i,1}}{s^t_{i,0}}\right)\middle| D_i = 0\right] - \mathrm{E}\left[\log\left(\frac{s^t_{i,1}}{s^t_{i,0}}\right)\middle| D_i = 1\right] = \theta\Big(\underbrace{\log \bar{q^t}}_{=\text{constant}} - E[\log v^t_{i,0}|D_i = 1]\Big)$$

This follows from the equilibrium condition for control users where $v^t_{i,1} = v^t_{i,0} = s^t_{i,0}$. First differences between treatment and control on the left-hand side account for unobserved heterogeneity in sharing behavior. □

C Details of Structural Estimation

This Appendix provides supplementary information on the structural model that enables estimation of some key parameters of interest.

C.1 Measurement Error

The design of the experiment may induce measurement error in the proportion of toxic content viewed. This is because of *sampling errors*, i.e. users view only a fraction of content in the ranked lists of content (in a set order), that the algorithm generates for them in each time period.

Among *treated users at baseline*, each toxic post *viewed* is assumed to be a Bernoulli trial with probability $q^t_{i,0}$. In each session therefore, the total number of toxic posts viewed is subject to measurement error, on account of the sampling procedure itself. However, since the sampling distribution of toxic views is known, the estimates can be corrected for measurement error using IV approaches (Schennach, 2016).

Consider the following linear classical measurement error set up. Suppose, $v^{t*}_{i,0}(1)$ denote the true proportion of toxic content viewed respectively, that are observed with measurement error in the data.

$$v^t_{i,0}(1) = v^{t*}_{i,0}(1) + ev^t_{i,0}(1)$$

where, $ev^t_{i,0}$ denote the measurement error in the proportion of toxic content viewed. In general, assume that $Cov(v^{t*}_{i,0}(1), ev^t_{i,0}(1)) = 0$. The estimators constructed from the strategy above are therefore, likely to suffer from attenuation bias due to the unobserved measurement error on the right-hand side of the estimating equation. I construct an instrumental variable to address this issue.

Note that $v^t_{i,0}$ is the average of toxic posts viewed over all the posts viewed (of any type)

by a user. Consider the proportion of toxic posts viewed out of half of the total posts viewed,

$$v_{i,0}^{\frac{t}{2(-)}}(1) = \frac{\sum_{j=1}^{J/2} t_{ij,0}(1)}{J/2}, \quad v_{i,0}^{\frac{t}{2(+)}}(1) = \frac{\sum_{j=1+J/2}^{J} t_{ij,0}(1)}{J/2}$$

where, $j \in \{1, \ldots, J\}$ indexes each post viewed by user i, so that $t_{ij,0}$ is a binary variable indicating whether post j was toxic or not, and $J/2$ indexes the median post. The first expression averages over the first half of posts per user (arranged in a random order) and is henceforth referred to as $half-1$. Similarly, $v_{i,0}^{\frac{t}{2(+)}}$ denotes the fraction of toxic posts out of the second half of the total posts viewed (for brevity, this variable is henceforth referred to as $half2$). However, assuming that the measurement errors pertaining to each half of the posts, are uncorrelated to each other for every user, this fraction computed over the first half of posts can be instrumented by this variable constructed using the second half of the posts. That is to say,

$$Cov(ev_{i,0}^{\frac{t}{2(-)}}(1), ev_{i,0}^{\frac{t}{2(+)}}(1)) = 0 \tag{AME}$$

Under this exclusion restriction, the attenuation bias in a 2SLS estimate of γ_1 is reduced to zero.

Proposition C.1. *Measurement error in average toxic views is corrected by instrumenting the fraction of toxic posts viewed in the first half of posts viewed ($half1$), with the fraction of toxic posts viewed in the second half of posts viewed ($half-2$) by a user in a session.*

Proof. The measurement error in these variables constructed using half the viewed posts, is written as

$$\begin{aligned} v_{i,0}^{\frac{t}{2(-)}}(1) &= v_{i,0}^{\frac{t*}{2(-)}}(1) + ev_{i,0}^{\frac{t}{2(-)}}(1) \\ v_{i,0}^{\frac{t}{2(+)}}(1) &= v_{i,0}^{\frac{t*}{2(+)}}(1) + ev_{i,0}^{\frac{t}{2(+)}}(1) \end{aligned}$$

where, as before $Cov(v_{i,0}^{\frac{t*}{2(-)}}(1), ev_{i,0}^{\frac{t}{2(-)}}(1)) = 0$ and $Cov(v_{i,0}^{\frac{t*}{2(+)}}(1), ev_{i,0}^{\frac{t}{2(+)}}(1)) = 0$.

Note the first stage regression using $half-2$ as the instrumental variable,

$$\begin{aligned} v_{i,0}^{\frac{t}{2(-)}}(1) &= \alpha_0 + \alpha_1 v_{i,0}^{\frac{t}{2(+)}}(1) + \mu_{i,0} \\ &= \alpha_0 + \alpha_1 (v_{i,0}^{\frac{t*}{2}}(+) + ev_{i,0}^{\frac{t}{2(+))}}(1) + \mu_{i,0} \end{aligned}$$

where, $Cov(v_{i,0}^{\frac{t}{2(+)}}(1), \mu_{i,0}) = 0$. Then, any bias in the estimates from the IV specification, due to measurement error in fraction of toxic posts viewed would depend on

$$\begin{aligned} Cov(v_{i,0}^{\frac{t}{2(-)}}(1), v_{i,0}^{\frac{t}{2(+)}}(1)) &= Cov(v_{i,0}^{\frac{t*}{2(-)}}(1) + ev_{i,0}^{\frac{t}{2(-)}}(1), v_{i,0}^{\frac{t*}{2(+)}}(1) + ev_{i,0}^{\frac{t}{2(+)}}(1)) \\ &= Cov(v_{i,0}^{\frac{t*}{2(-)}}(1), v_{i,0}^{\frac{t*}{2(+)}}(1)) \end{aligned}$$

Therefore, the IV approach eliminates measurement error, due to the exclusion restriction stated in (AME). This shows that the IV estimation strategy only depends on the true

distribution of the main explanatory variable. □

C.2 Validation of structural estimates

I validate my estimation procedure that measures the rate at which users update their sharing behavior upon being randomly exposed to more non-toxic content during the intervention period. This model correctly estimates the updating-behavior only for treated users, because for these users, exposure to toxic content in the baseline period is related to the engagement with such content only through the channel of behavioral response.

θ is not estimable using the sample of control users as these users are always in equilibrium, meaning that viewing and sharing behaviors are perfectly correlated. Therefore, estimates that employ the control sample are expected to be distinct from the main estimates derived using data on treated users only. Figure D.19c shows that this is indeed the case. Additionally, Figure D.19b shows that exposure to toxic content *during the intervention period* has a much smaller effect on the odds of sharing such content. This also validates the main result, because the intervention period exposure is very likely concentrated around the average user's exposure, and is expected to produce different estimates.

C.3 Calibration

I match moments of the empirical distribution of various outcomes, with the distributions simulated by the model using $\hat{\theta} = 0.16$. This enables calibration of four main parameters of the model: (1) β, the consumption value of viewing posts, (2) α, the disutility from viewing unshareable posts, (3) η, the cost of sharing an additional post, (4) δ, the utility weight on conformity with societal norms. I use the method of simulated moments to estimate these parameters, using the data $\{s^t_{i,1}, v^t_{i,1}, S_{i,1}, N_{i,1}\}$, which is the proportion of toxic posts shared and viewed respectively, as well as the number of posts shared and viewed, respectively. I compute the empirical mean of each of these outcomes, separately for users with above and below median exposure to toxic content at baseline.

$$\mathrm{E}[X] = \frac{1}{n/2}\sum_{i=1}^{n} x_i$$

Then, the model is defined by the following functions using the equilibrium conditions,

$$s^t(v^t, p^t, \theta) = \left(v^t\right)^{\theta}\left(p^t\right)^{1-\theta} \tag{23}$$

$$N(\delta, \theta, \alpha, \eta, \beta, v^t, p^t) = \frac{-\alpha\delta\theta\left(\log\frac{v^t}{p^t}\right)^2 + \beta(\alpha+\eta)}{2\alpha\eta} \tag{24}$$

$$S(\delta, \theta, \alpha, \eta, N, v^t, p^t) = \frac{\delta\theta(1-\theta)\left[\left(\log p^t\right)^2 - 2\log v^t \log p^t + \left(v^t\right)^2\right] + \frac{N}{\eta}}{2(\alpha+\eta)} \tag{25}$$

where, v^t, the proportion of viewed posts that are toxic. The moment conditions for users with lower proclivity to toxic content are given as,

$$\mathrm{E}_1[s^t] = \int_0^m \left(v^t\right)^\theta \left(p^t\right)^{1-\theta} \cdot f(v^t) dv^t \tag{26}$$

$$\mathrm{E}_1[N] = \int_0^m N(\delta, \theta, \alpha, \eta, \beta, v^t, p^t) \cdot f(v^t) dv^t \tag{27}$$

$$\mathrm{E}_1[S] = \int_0^m S(\delta, \theta, \alpha, \eta, N, v^t, p^t) \cdot f(v^t) dv^t \tag{28}$$

where, m denotes the median value of the proportion of toxic posts shared, v^t, at baseline. Similarly, I write the moment conditions for users with higher proclivity to toxic content as,

$$\mathrm{E}[s^t] = \int_m^\infty \left(v^t\right)^\theta \left(p^t\right)^{1-\theta} \cdot f(v^t) dv^t \tag{29}$$

$$\mathrm{E}[N] = \int_m^\infty N(\delta, \theta, \alpha, \eta, \beta, v^t, p^t) \cdot f(v^t) dv^t \tag{30}$$

$$\mathrm{E}[S] = \int_m^\infty S(\delta, \theta, \alpha, \eta, N, v^t, p^t) \cdot f(v^t) dv^t \tag{31}$$

I use numerical integration methods to evaluate these integrals, assuming $v^t \sim EVT1$. Subsequently, the empirical moments are matched with the simulated moments. I use the Nelder-Mead simplex method to estimate the parameters of the model, which converge to the following values in 800 iterations, in this case (Gao and Han, 2012).

C.4 Discussion

I use a structural model to formalize the analysis for the following reasons. First, the model's equilibrium characterization of user types allows an analysis of the treatment effect on toxic sharing. This is because user preferences are not observed, but are inferred from the assignment probabilities of toxic posts at baseline, when the system is assumed to be in equilibrium. Second, the model provides micro-foundations for user engagement with harmful content. In the model, users update their view of socially acceptable content in order to conform with other users of similar type (Fang and Loury, 2005). Treated users were served an average user's feed, and thought to update their opinion of what other users of the same type might be viewing.

Third, the model decomposes the treatment effect into two channels with various counterfactual policies that cannot be implemented in the data. This includes **(1)** the endogenous response in the total number of posts viewed and shared, and **(2)** the influence of exposure to diverse content on the proportion of toxic posts shared. Treated users endogenously responded to diversity in content assignment by viewing fewer posts, or spending less time on the platform. However, the model shows that this effect was more pronounced for users who were previously engaged with more extreme toxic content (henceforth, toxic users). This provides valuable information to a regulator interested in policies that reduce the total amount of toxic content shared on social media platforms, to compare the costs and benefits of such

an intervention. Finally, the structural model estimates the malleability of user behavior by identifying θ. However, the model makes a set of assumptions that are validated below.

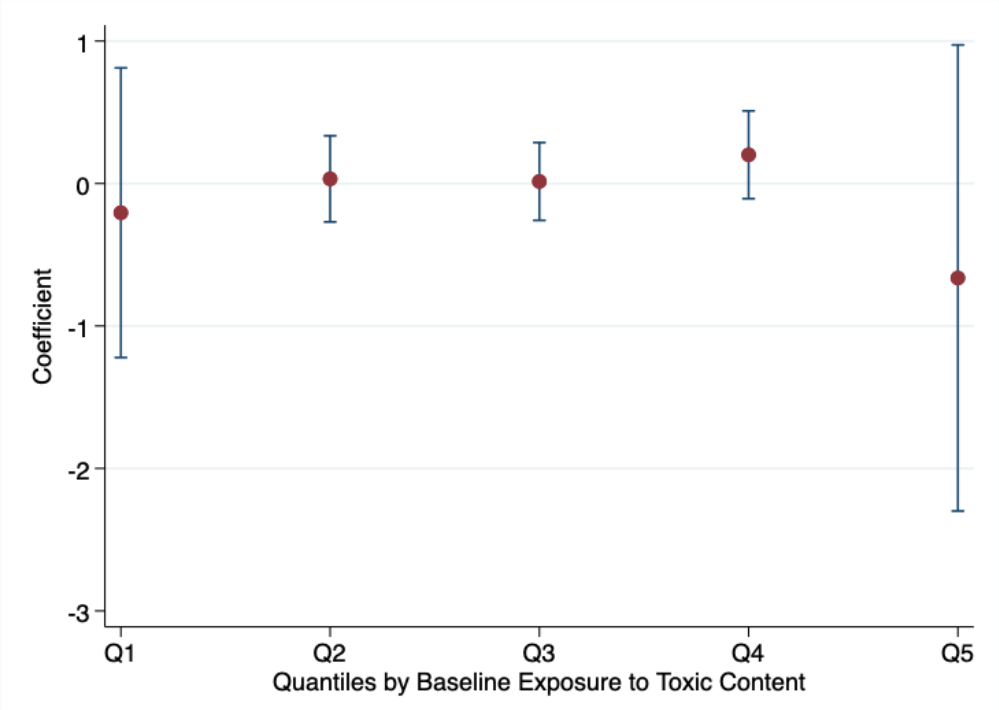
C.4.1 Simplifying Assumptions

In writing the utility for structural estimation, I made four simplifying assumptions: **(1)** consumption as well as signalling utilities are additively separable for each content type, **(2)** user behavior in the action-signalling model is updated at some constant rate θ across all users, **(3)** deviating from the reference point of own and society's tastes generates disutility which is quadratic in nature.

The first assumption rules out strategic complementarities and substitutabilities between different kinds of posts. This is tenable due to the fact that users scrolling through social media are assumed to be viewing posts one at a time, and do not know if the next post they will view is going to be toxic or not.

I test the second assumption, that is, the constant effects with respect to the rate of updating user behavior in the action-signalling model. Figure 9 supports this assumption, as the estimates of θ obtained from samples of different types of users are indistinguishable from each other. Finally, I have assumed the costs of using social media to be quadratic for ease of computation. The model does not stray far from the literature on strategic interactions in the presence of social signalling, especially when such models are estimated using structural methods, for instance, in Butera et al. (2022).

Figure 9: Testing simplifying assumption in action-signalling model



Notes: This Figure shows that I cannot reject the hypothesis that heterogeneous users are influenced by exposure at equal rates. The plot was obtained by estimating the main structural equation from the model, in different sub-samples of users, based on their baseline toxic exposure.

C.4.2 Identifying Assumptions

The main identifying assumption in this framework is that the probability of sharing toxic content is equal in steady state equilibrium. Since, the control users remain in steady state during the intervention and were chosen randomly, I test this assumption in the sample of control users,

$$s^t_{i,0} = s^t_{i,1} \tag{IA}$$

and estimate parameters of the following regression using normalized proportion of toxic content shared in each time period,

$$s^t_{i,1}(D_i = 0) = \delta_1 s^t_{i,0}(D_i = 0) + \varepsilon_{i,1}$$

Under this identifying assumption (IA) I expect $\delta_1 = 1$ in the sample of control users. Table 4 shows that I cannot reject the hypothesis that $\delta_1 = 1$, in the measurement error corrected case.

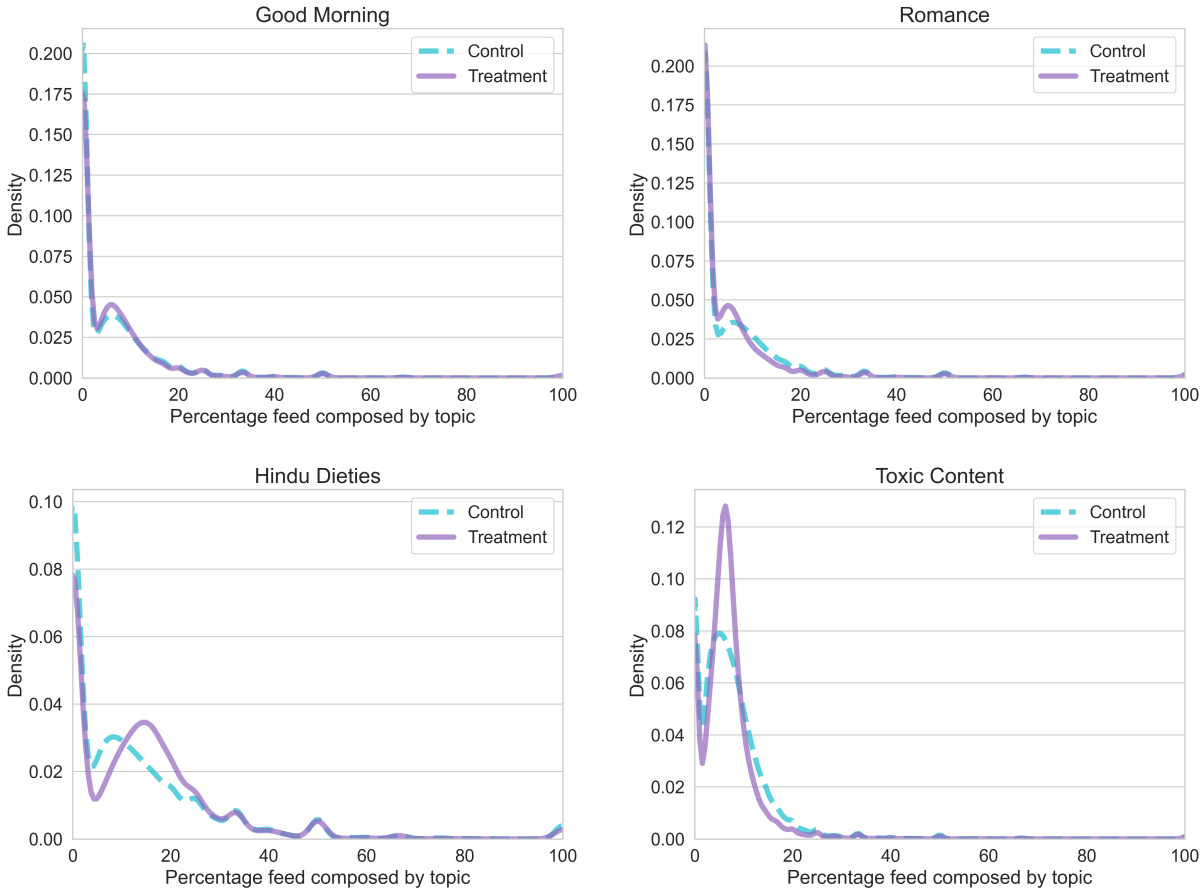
Table 4: Testing identifying assumption in structural model using control sample

	(1)	(2)	(3)
	Probability of sharing toxic post during intervention period		
Proportion of toxic posts shared at baseline	0.112^{***}		0.820^{***}
	(0.012)		(0.091)
Proportion of toxic posts among first half of posts shared at baseline		0.290^{***}	
		(0.057)	
N	52663	52663	52663

Notes: This Table tests the identifying assumption, derived from the steady state condition $s^t_{i,0} = s^t_{i,1}$. That is, all else equal, the probability of sharing toxic content for each user is expected to be equal in each time period. Column (3) shows that the measurement error corrected estimates of the slope coefficient is close to 1. The sample includes control users who shared at least one post at baseline. Robust standard errors in parentheses. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$

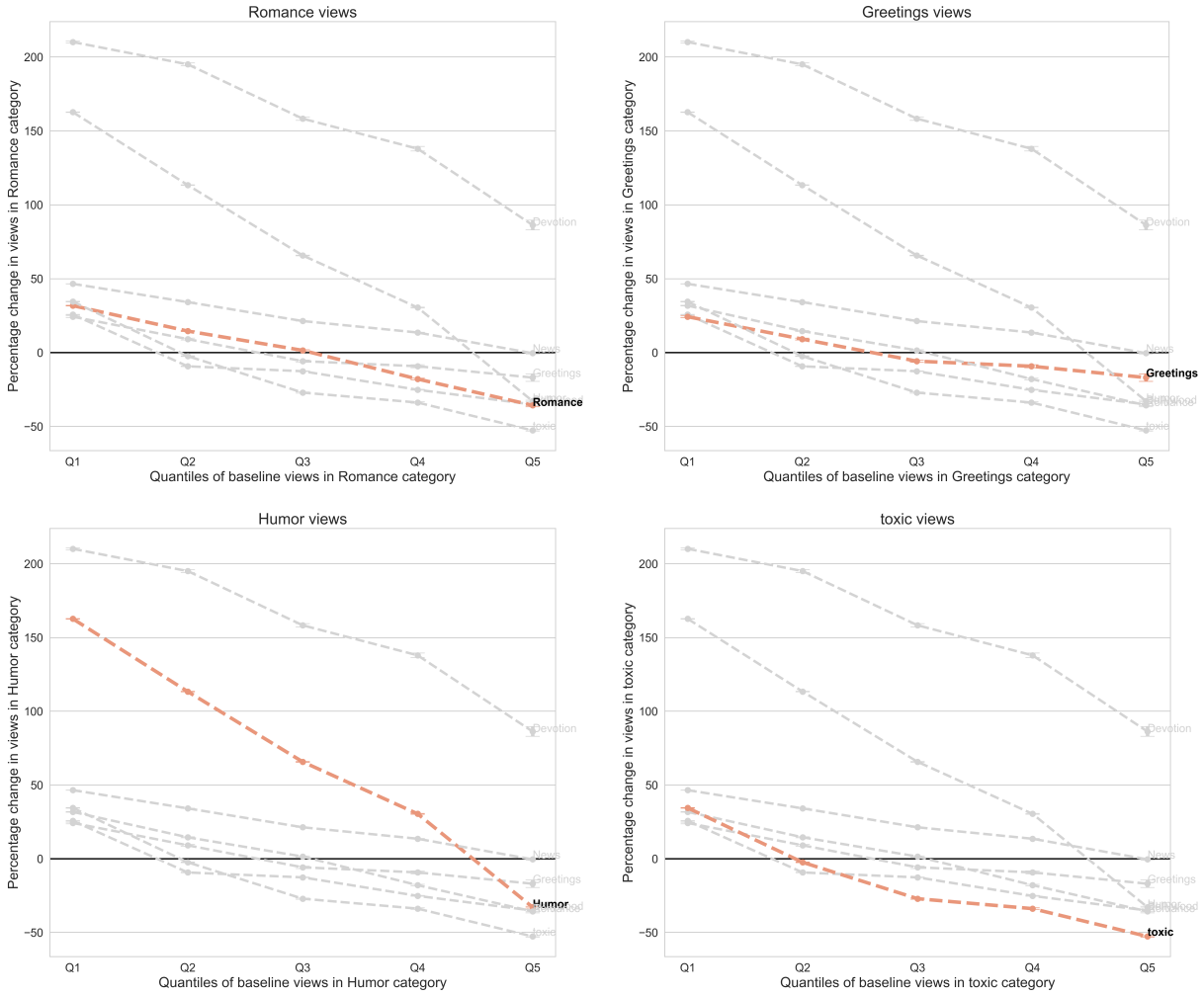
D Supplementary Figures

Figure D.1: Change in distribution of exposure to different topics



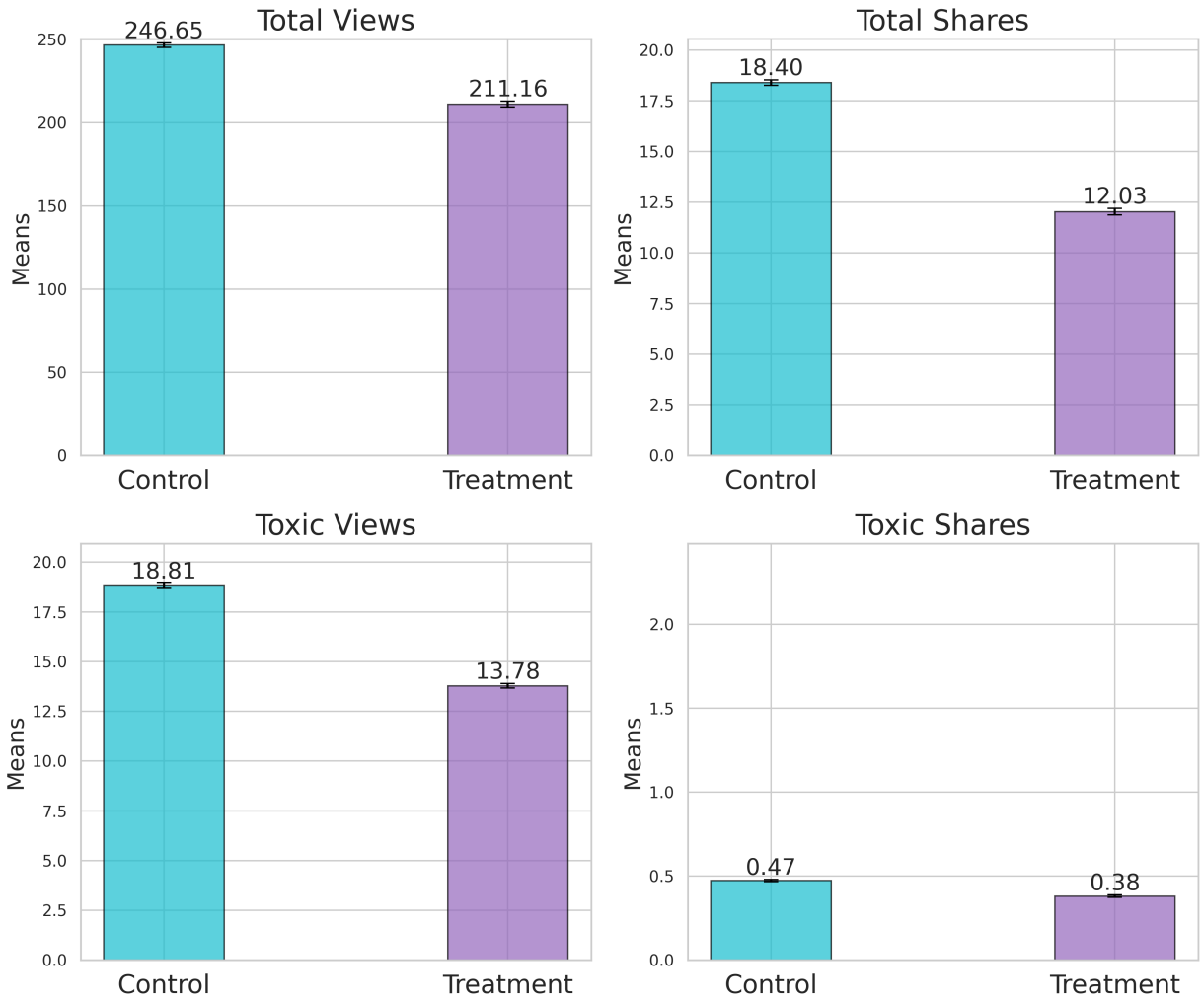
Notes: This Figure shows the change in the distribution of exposure to different types of content for treated users. The top two panels show minimal change in the distribution of exposure to good morning posts and romantic content. The bottom two panels show the change in the distribution of exposure to toxic content, as well as religious posts on various Hindu deities (one for each day of the week). The distribution of toxic and religious content is more concentrated and shifted to the left for toxic content, and shifted to the right for religious content. The topics were modeled using an LDA topic model, and the dominant topic in each post was computed. Post topic and toxicity are not mutually exclusive.

Figure D.2: Change in distribution of exposure to different types of content



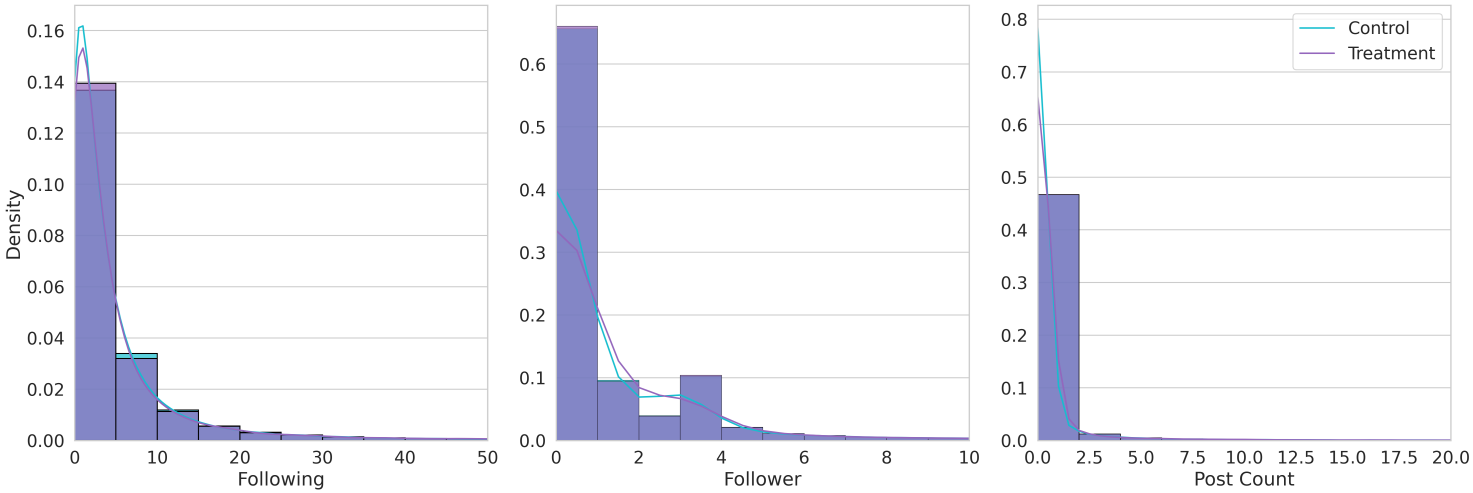
Notes: This Figure shows that among all broad content categories, the percent reduction in exposure to content is largest for toxic posts. This provides an experiment to measure the effect of exogenously reducing exposure to harmful content on user behavior. In each plot, the quantiles represent the percentage of feed that is composed of the given content type at baseline. Therefore, Q5 of the top left plot consists of users who saw the most romantic content at baseline. Post toxicity and content categories are not mutually exclusive.

Figure D.3: Comparison of means across treatment and control, for key outcomes



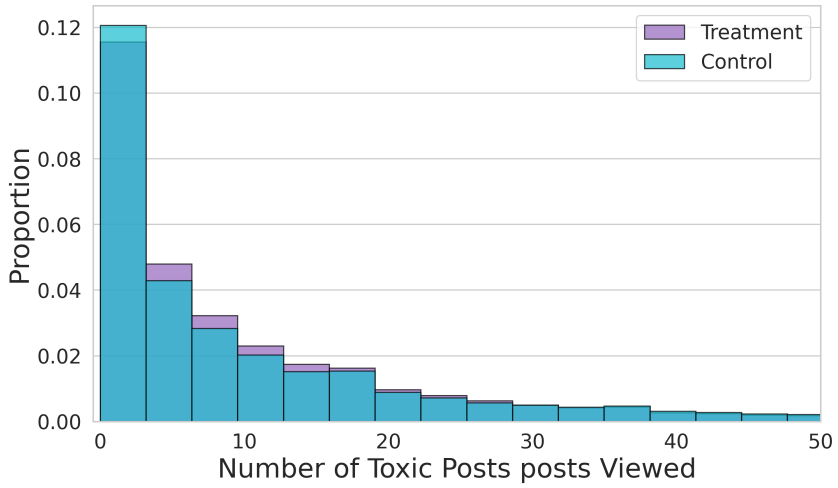
Notes: This Figure shows the trade-off between user engagement and the propagation of harmful content on social media. During the first month of the intervention, treated users were, on average, exposed to less toxic content, but were also less active on the platform. This highlights the costs (in terms of reduced user engagement with the platform), and benefits (in terms of reduced engagement with toxic content) of the intervention. Further, the decrease in toxic shares is not as large as the total decrease in shares, or the decrease in toxic views. User behavior is said to be inelastic with respect to toxic content because the ratio toxic shares to total shares is significantly higher in the treatment (3.16%), than in the control group (2.55%). Stickiness in sharing behavior with respect to toxic content is explained by the structural model, which quantifies the extent to which reduced toxic sharing is driven by the influence of reduced exposure to toxic content. Standard errors are depicted using confidence intervals around the means.

Figure D.4: Distinct features of SM's interface

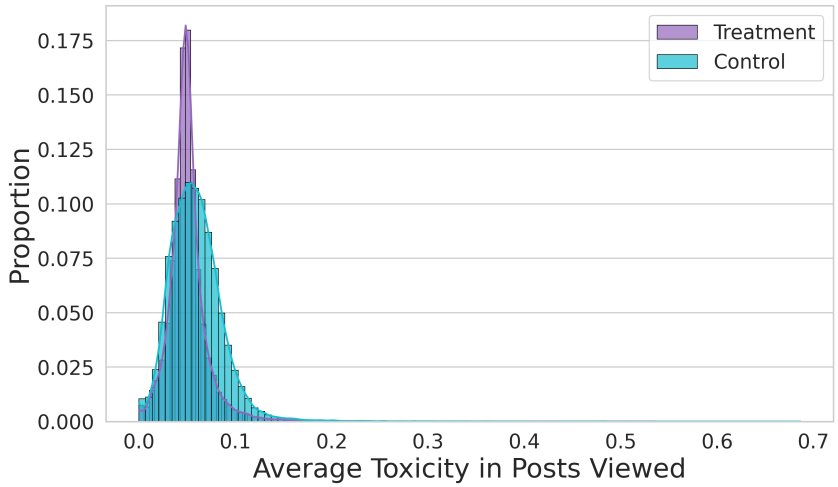


Notes: This figure shows that SM's user interface is distinct from other social media platforms. In particular, despite having an option to follow other users, the platform is content-based, and users interact with content rather than with other users. This is in contrast to platforms like X (formerly, Twitter), where users engage with users they 'follow.' Additionally, SM is distinct from other platforms because users consume content produced by influencers (and not by friends and family), and very rarely produce their own content.

Figure D.5: Distribution of exposure to toxic content during intervention period



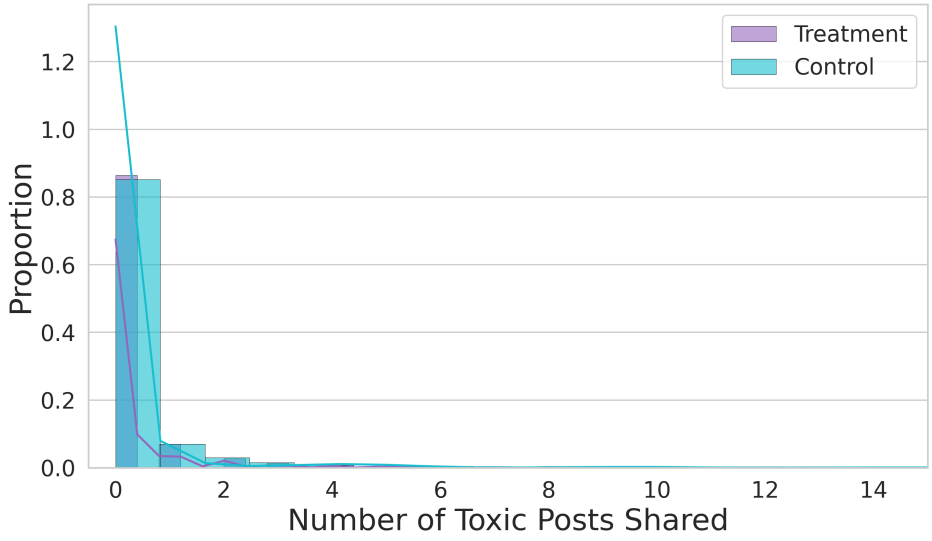
(a) Number of toxic posts viewed (with binary indicator for post toxicity)



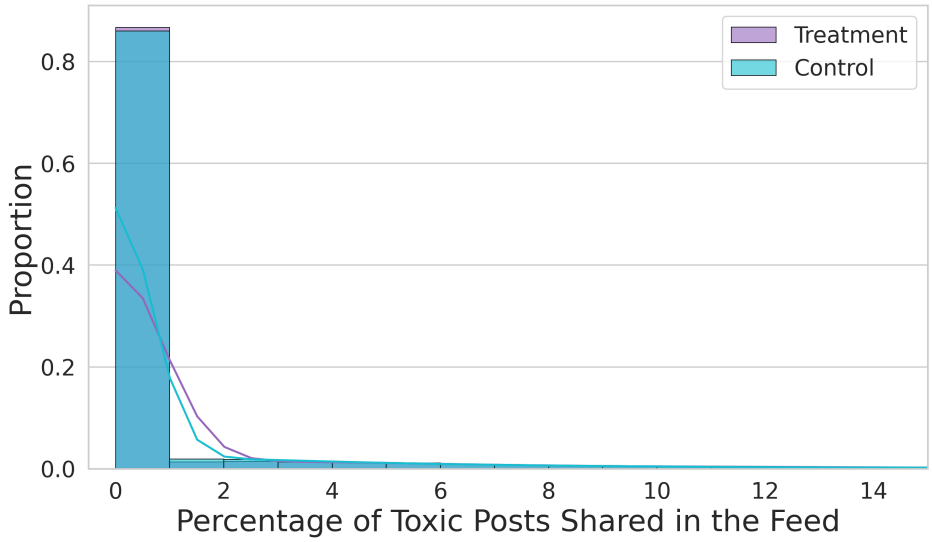
(b) Average toxicity scores on posts viewed (with continuous toxicity scores)

Notes: This Figure plots the raw data on the number of toxic posts viewed by users during the intervention period. The top panel uses the 0.2 threshold to classify a post as toxic, which generates a binary variable. The bottom panel uses the continuous toxicity score to measure the average toxicity of a user's feed. The distribution of toxic views for control users is to the left of the distribution for treated users. This is consistent with the main result that the intervention reduced exposure to toxic content for the average user in the treated user in the experimental sample.

Figure D.6: Distribution of engagement with toxic content during intervention period



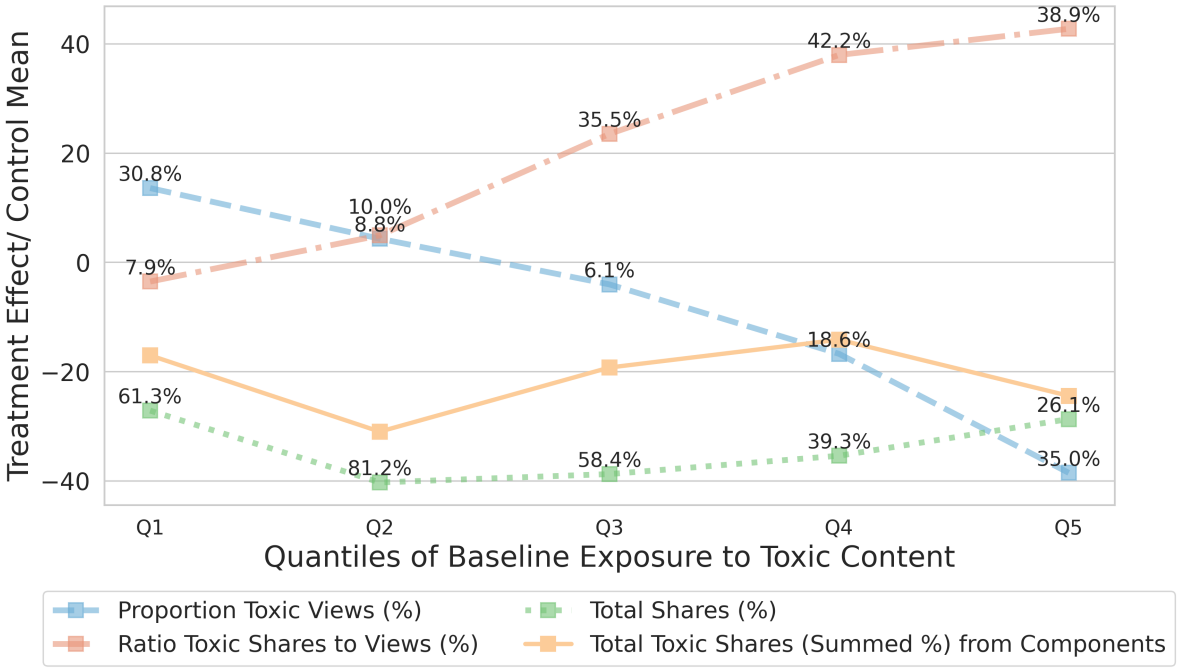
(a) Number of toxic posts shared



(b) Percentage of posts shared that are toxic

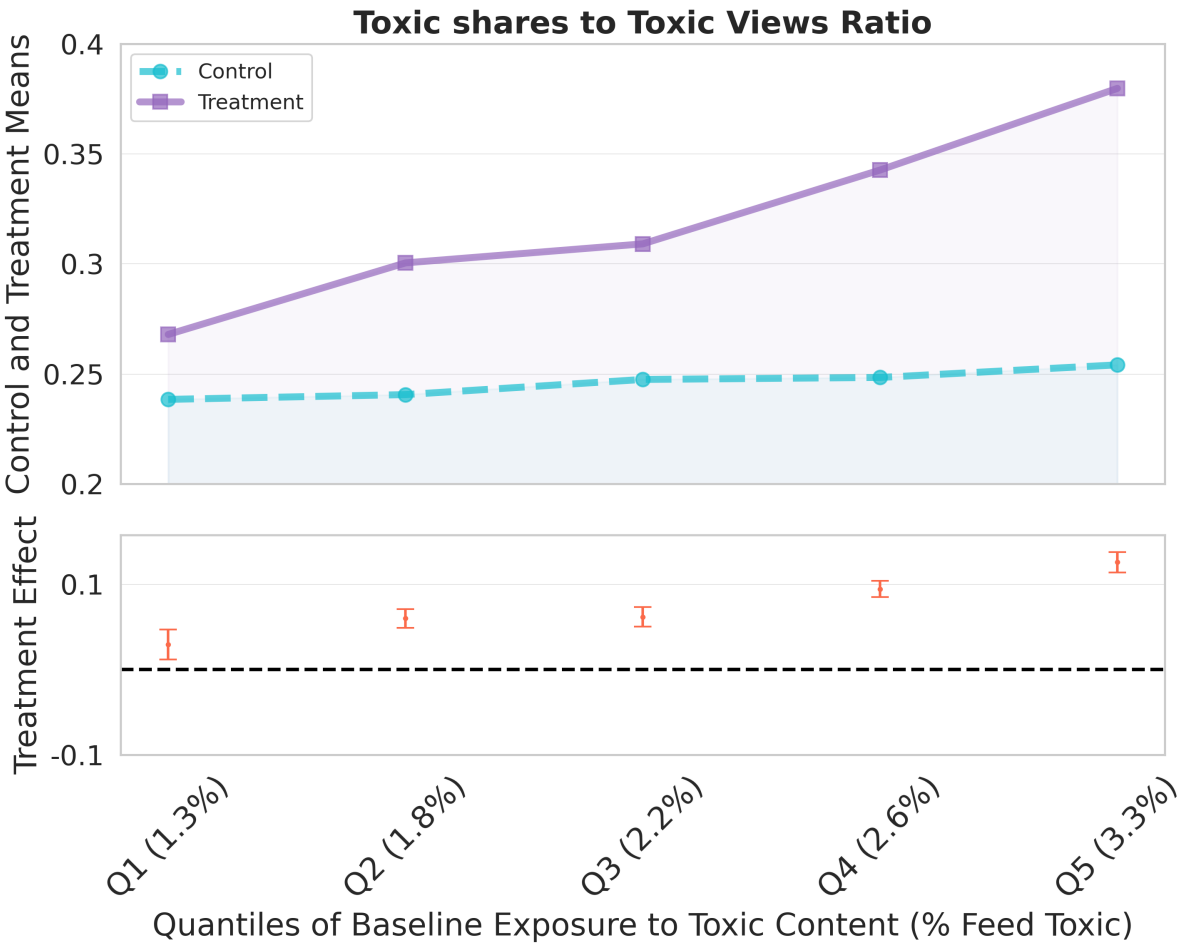
Notes: This Figure plots the raw data on toxic shares for treated and control users, and shows that the distribution toxic posts shared by control users first order stochastically dominates the distribution for treated users. Panel (a) provides the number of toxic posts shared, where a toxic share is defined as a shared post with toxicity score greater than 0.2. Panel (b) provides the percentage of shares that are toxic, where the proportion is defined as the number of toxic shares divided by the total number of shares.

Figure D.7: Empirical decomposition of treatment effects on toxic shares



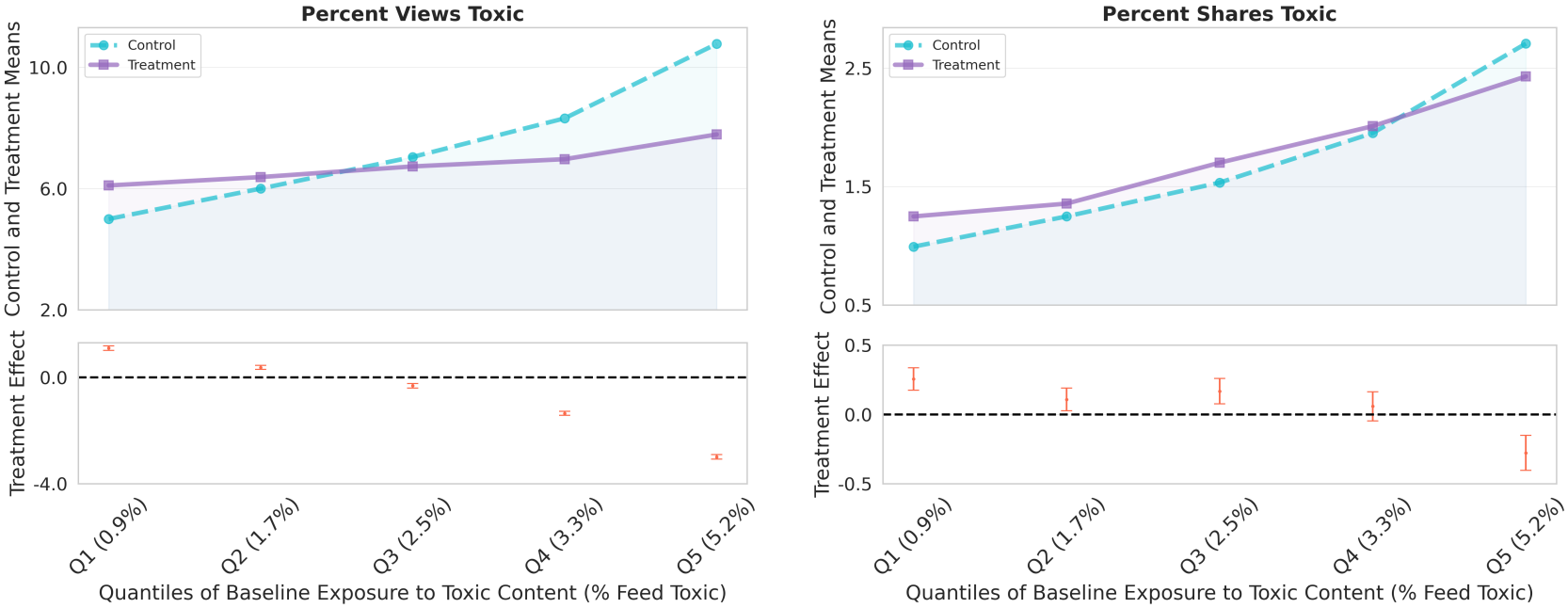
Notes: This Figure shows replicates the main result, that the ratio of toxic shares to toxic views is increasing in user's baseline toxicity, even when toxicity is measured by averaging over the continuous toxicity scores of posts viewed or shared. On average, the ratio of toxic shares to toxic views is higher among the treated, and this is driven by users with higher proclivity to toxic content. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.8: Change in sharing relative to toxic views for continuous toxicity scores



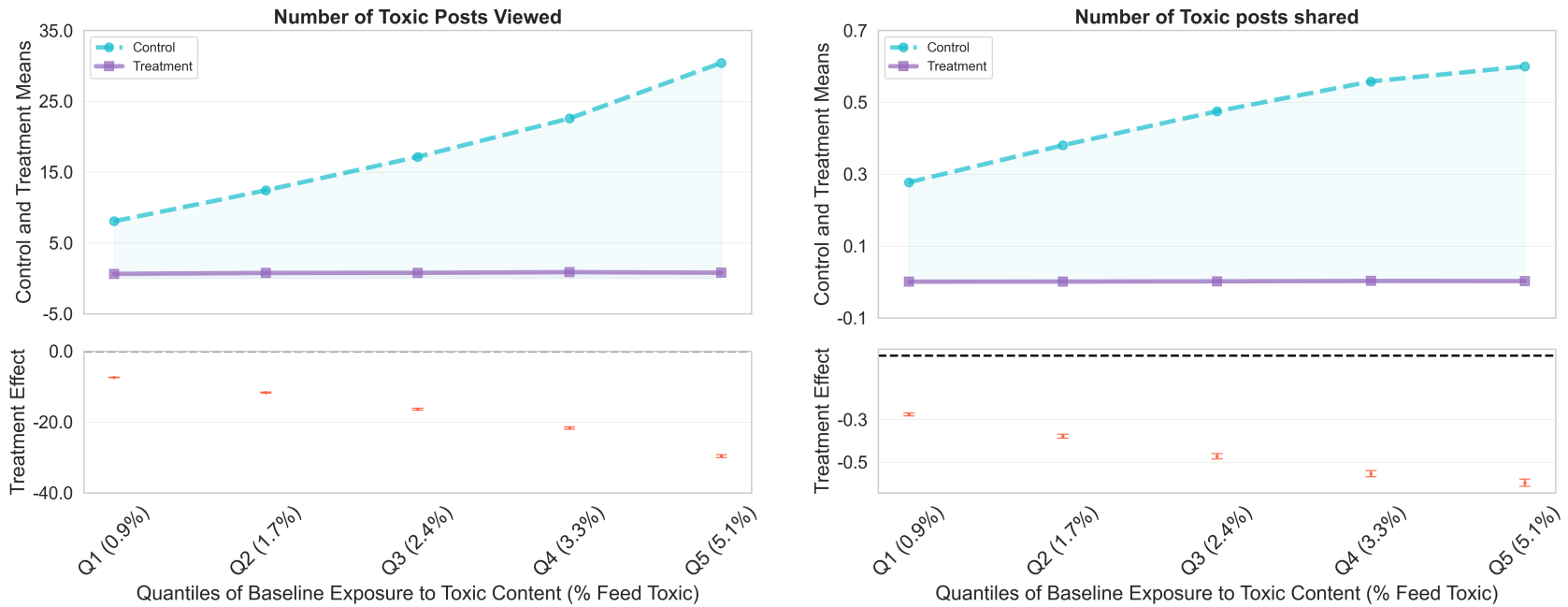
Notes: This Figure shows replicates the main result, that the ratio of toxic shares to toxic views is increasing in user's baseline toxicity, even when toxicity is measured by averaging over the continuous toxicity scores of posts viewed or shared. On average, the ratio of toxic shares to toxic views is higher among the treated, and this is driven by users with higher proclivity to toxic content. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.9: Treatment effects on toxic behavior as percentage of total engagement (views and shares)



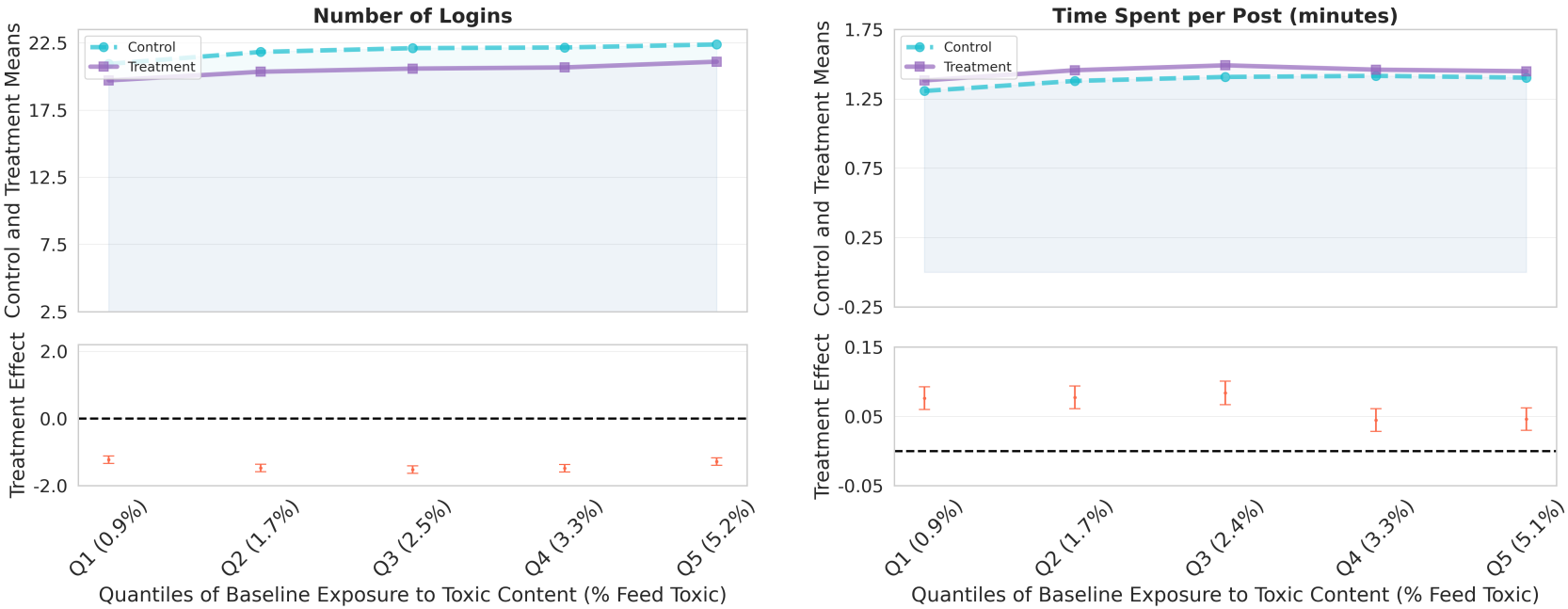
Notes: This figure shows that the treatment effect, on the *proportion* of posts shared that are toxic, is *non-negative* for all users except those in Q5 (with the highest exposure to toxic content at baseline). This is true, even in the cases of Q3 to Q5, where user type is toxic enough that the treatment effect on toxic views is negative (left panel). The model predicts positive treatment effect on the proportion of toxic shares for users with lower degree of proclivity to toxic content, but decreased overall engagement with the platform from more toxic users. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.10: Treatment effects on toxic behavior with respect to political posts only



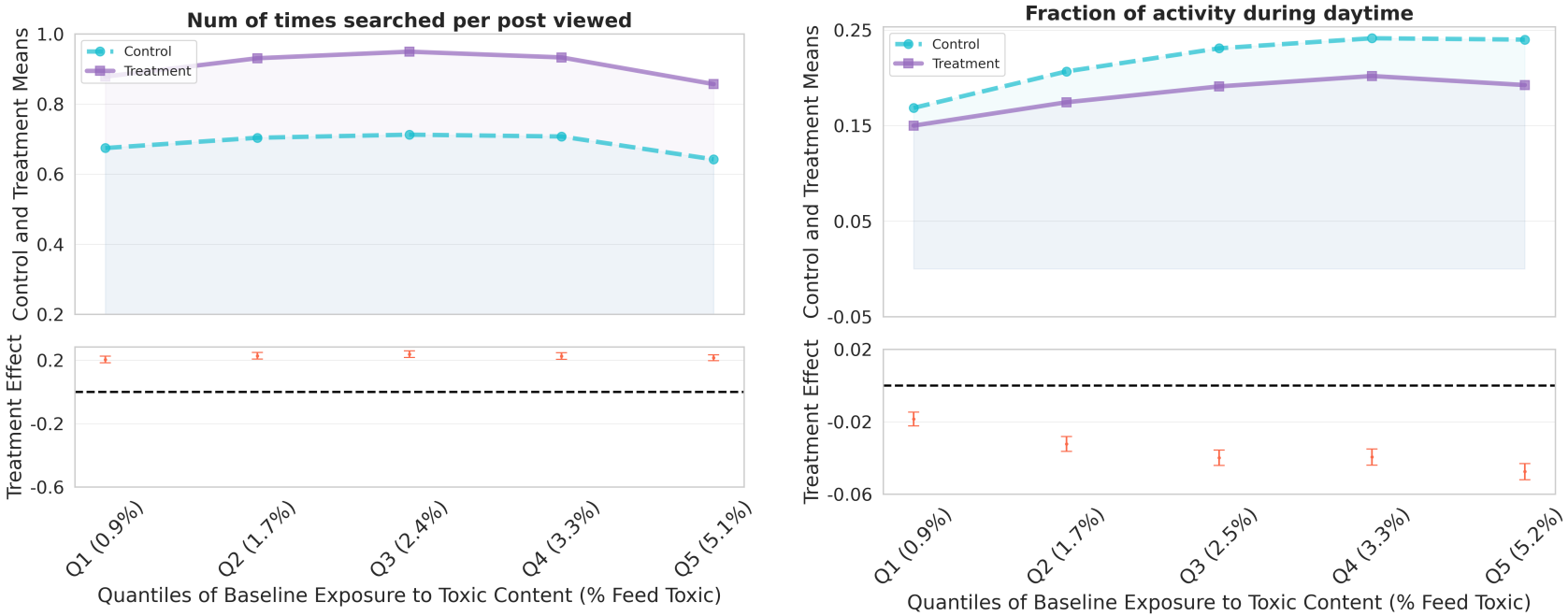
Notes: This figure shows that the treatment effect on the number of toxic posts viewed and shared is negative in the sample of political posts only. This is true, even in the cases of Q3 to Q5, where user type is toxic enough that the treatment effect on toxic views is negative (left panel). The model predicts positive treatment effect on the proportion of toxic shares for users with lower degree of proclivity to toxic content, but decreased overall engagement with the platform from more toxic users. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.11: Treatment effects on platform activity, by user type



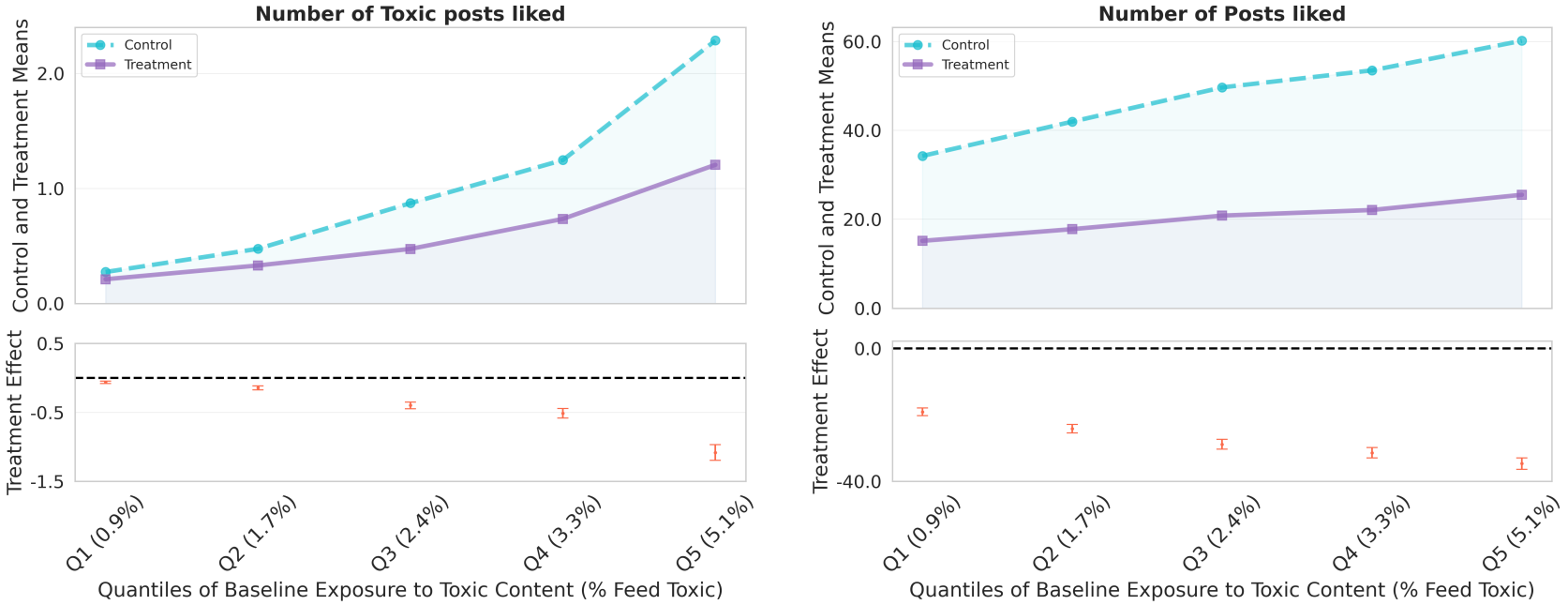
Notes: This figure shows that the treatment effect, on the overall activity of users on the platform is negative for all users. There is a decrease in the number of times a user logged in. The effects on logins are indistinguishable across user types. There is an increase in time spent per post, but the increase is much smaller for more toxic users (Q4 and Q5). This means such users are more likely to scroll through posts, and spend less time on each post. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.12: Auxiliary evidence of users seeking out preferred content



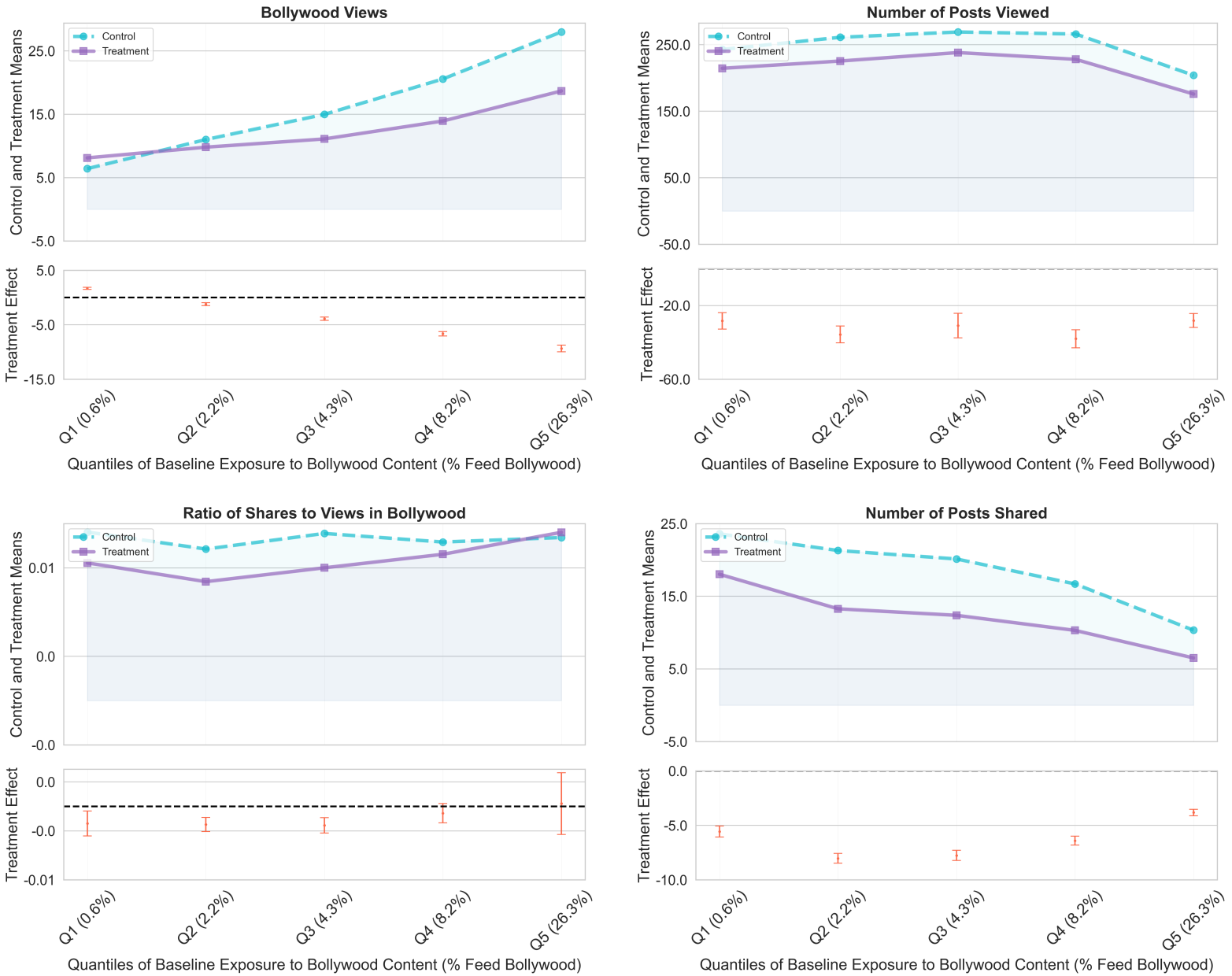
Notes: This figure shows that the intervention changed the quality of user engagement on the platform. This is because treated users were more likely to use the platform during the weekend, or in the night. Users spent less on the platform during their working hours. This has significant implications for the labor-leisure trade off, and questions about digital addiction, that are explored in a companion paper. The bottom panel in each figure shows the magnitude of treatment effects (coefficient plots), and the top panel shows the means of the outcome variable in each quantile of treatment and the control group. All regressions were run at the user level, and robust standard errors were computed.

Figure D.13: Experimental effects on liking behavior, by user type



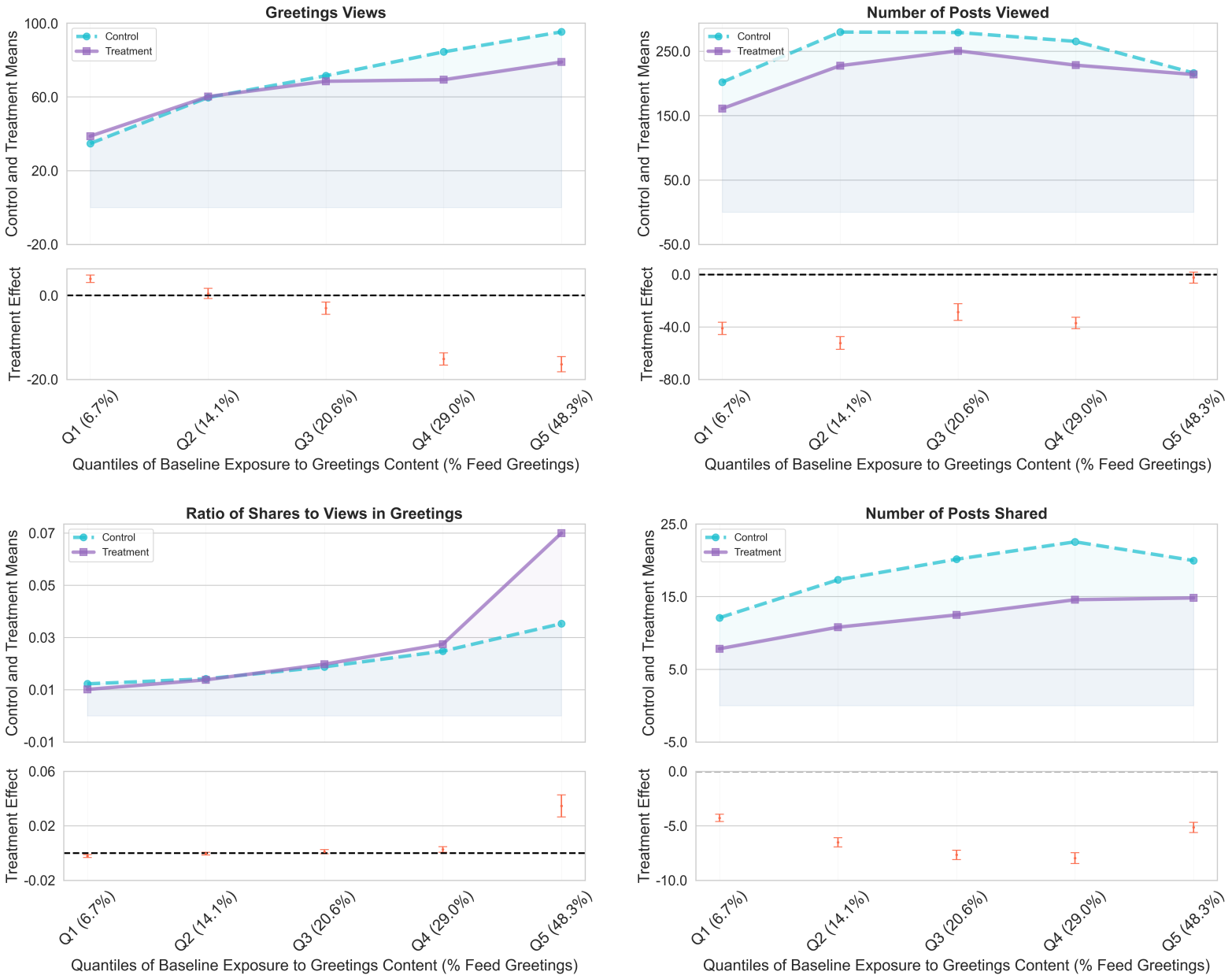
Notes: This figure shows that the treatment effect, on the *proportion* of posts shared that are toxic, is *non-negative* for all users except those in Q5 (with the highest exposure to toxic content at baseline). This is true, even in the cases of Q3 to Q5, where user type is toxic enough that the treatment effect on toxic views is negative (left panel). The model predicts positive treatment effect on the proportion of toxic shares for users with lower degree of proclivity to toxic content, but decreased overall engagement with the platform from more toxic users. On average, the ratio of toxic shares to toxic views is higher among the treated, and this is likely driven by users with low to medium proclivity to toxic content. The axis corresponding to the bottom plots show the magnitude of treatment effects (as coefficient plots), while the top panel is scaled according to the control mean of the outcomes for each quantile. All regressions were run at the user level, and inference about the treatment effects is based on robust standard errors.

Figure D.14: Treatment intensity with respect to Bollywood content



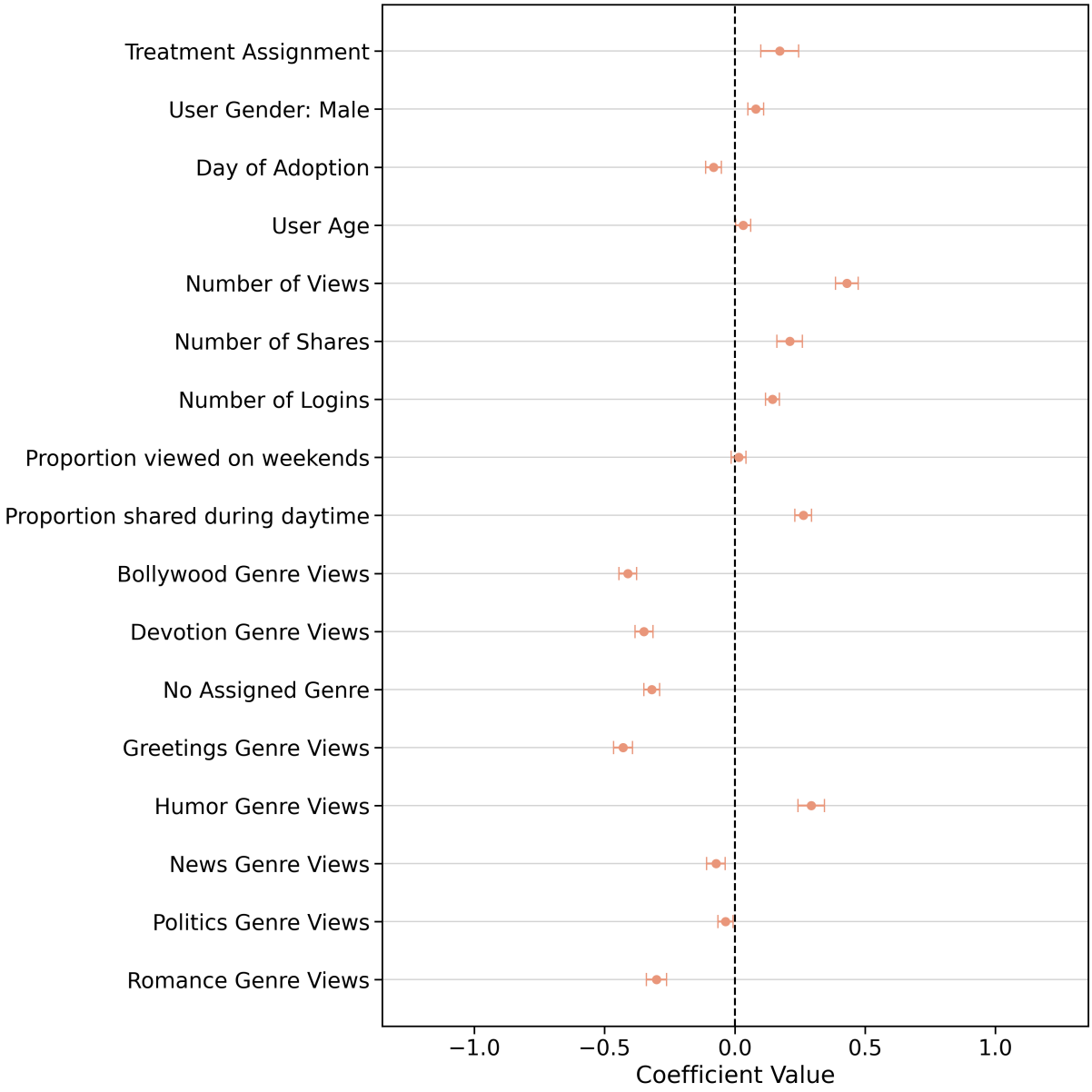
Notes: This Figure shows that the treatment effects on the number of Bollywood related posts mimics the treatment intensity with respect to toxic content. However, the treatment effects on the total number of posts viewed (of any type) do not follow the same pattern, according to user type defined in terms of proclivity to Bollywood content. This suggests that users do not seek out this type of content, and presumably access it on other platforms. Here, users are divided into quantiles based on their exposure to Bollywood content at baseline, which is a proxy for their proclivity to such content.

Figure D.15: Treatment intensity with respect to Greetings content



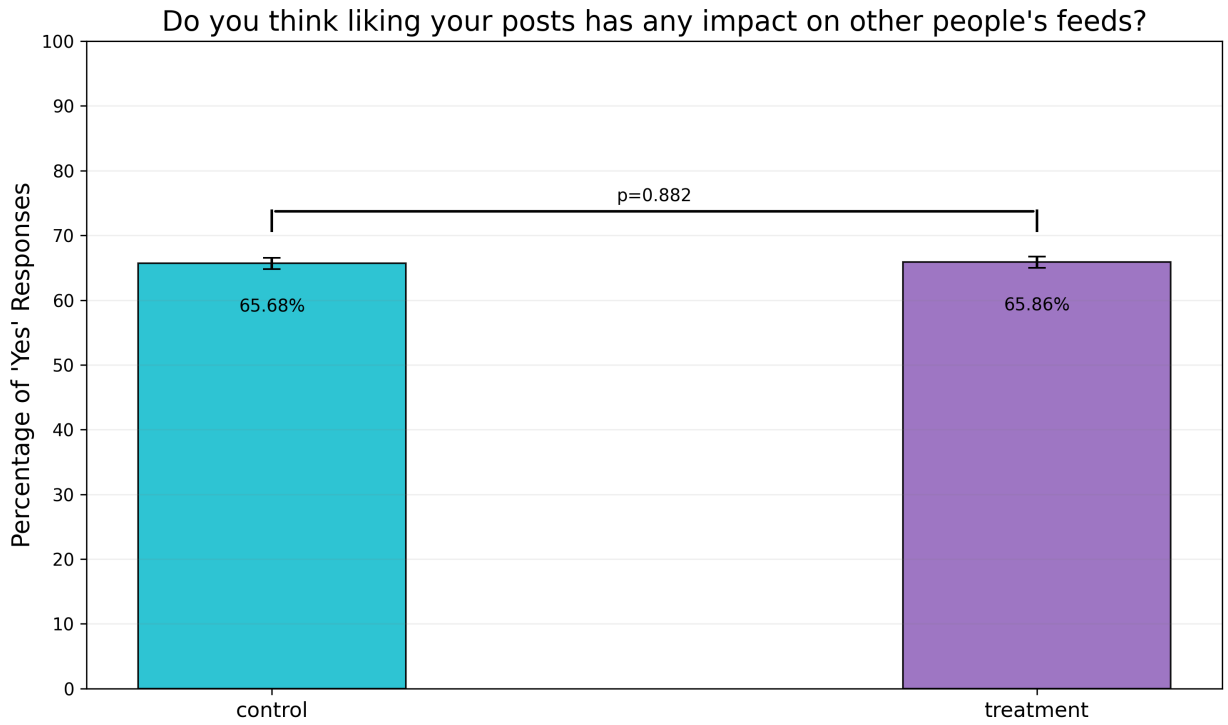
Notes: This Figure shows that the treatment effects on the number of Greetings related posts mimics the treatment intensity with respect to toxic content. The effects on the number of posts viewed in this genre, and the ratio of shares to views in the Greetings category, follow patterns similar to those observed for toxic content. However, the number of posts viewed (of any type) do not follow the same pattern as before. This is consistent with the explanation that users seek out content that they like, especially when Greetings content is not available on other platforms in India. Users are divided into quantiles based on their exposure to Greetings content at baseline, which is a proxy for their proclivity to such content.

Figure D.16: Suggested mechanisms driving engagement with toxic content



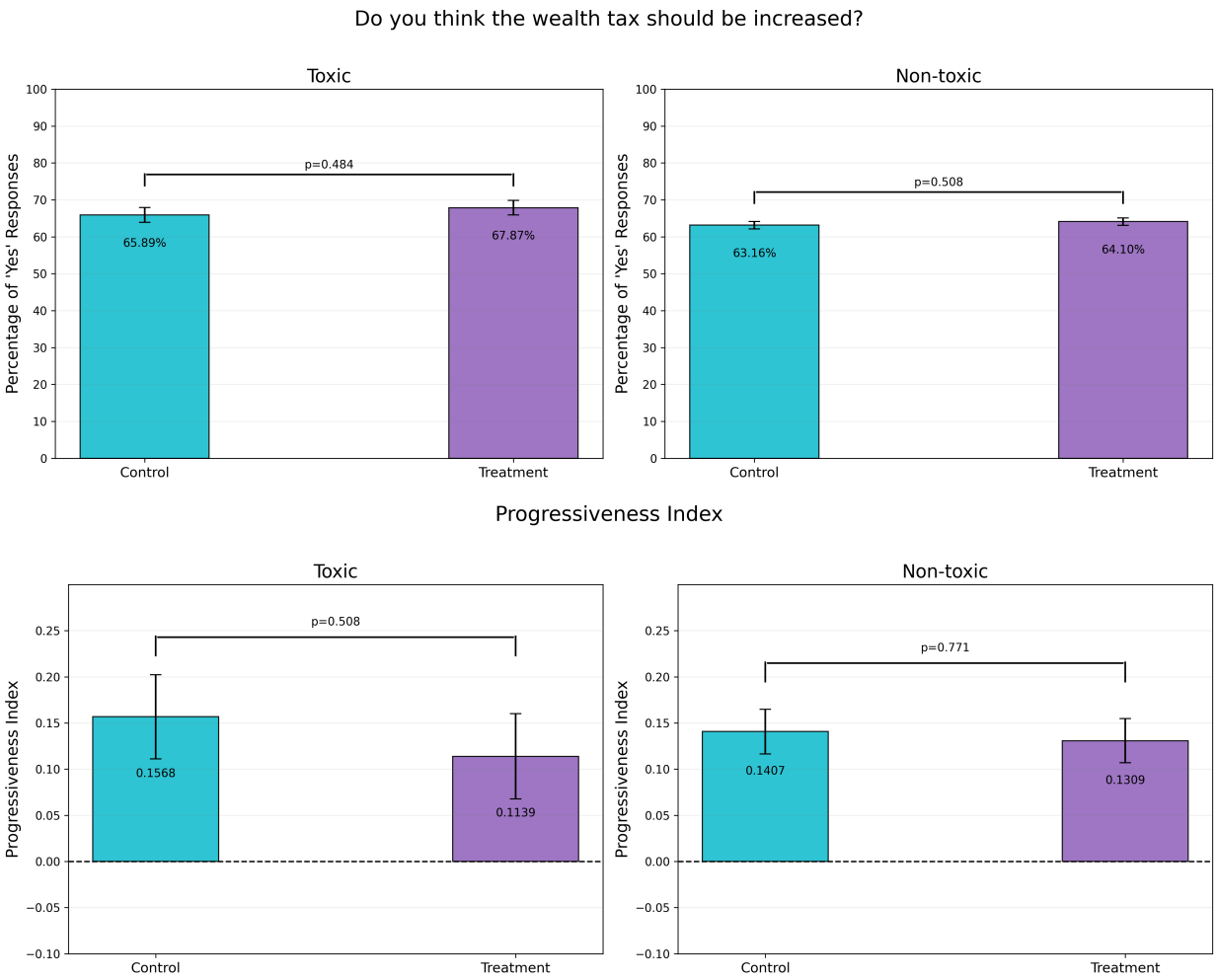
Notes: This Figure shows suggestive evidence on the mechanisms driving the treatment effects, by regressing the main outcome variable (proportion of shares that are toxic), treatment intensity (proportion of views that are toxic), as well as baseline user characteristics in the sample of treated users. For all types of users, toxic sharing is positively correlated with treatment intensity, and this correlation was shown to be the strongest for users with high proclivity to toxic content. Higher platform activity at baseline is associated with higher toxic sharing during intervention period, for all types of users. All variables were standardized as z-scores to get comparable magnitudes.

Figure D.17: Salience of personalization algorithms



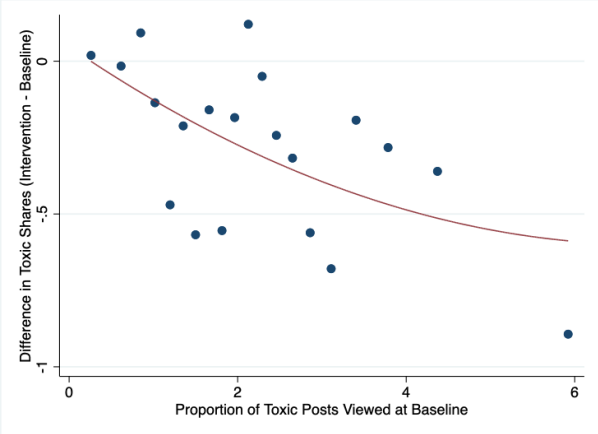
Notes: This Figure shows that the personalization algorithm is salient to users. A subset of users in the experimental sample were randomly selected for a follow-up survey ($N = 8,387$), and asked whether they thought their likes and shares changed the content in other users' feeds. More than 65% of the users said that they believed that their SM activity changes other people's feeds, and there were no differences in this response by treatment status. Uncertain responses were dropped before computing these percentages, and the error bars report standard errors of the means. The survey was conducted at the end of the intervention period, with $4,236$ users randomly sampled from the treatment group, and the remaining $4,151$ users sampled from the control group.

Figure D.18: Preferences over redistribution, by user type

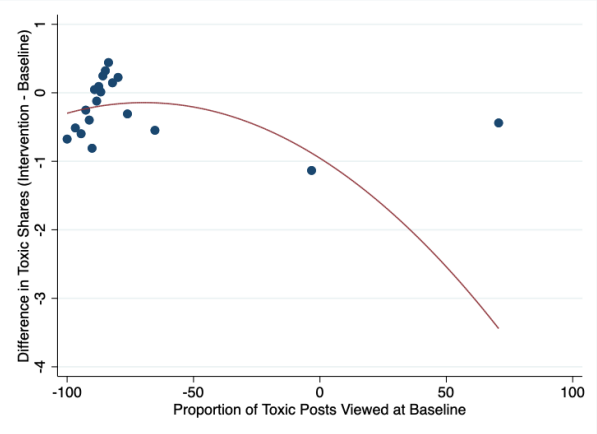


Notes: This Figure shows that the treatment did not affect users' preferences over redistribution, as reflected in the survey data (N = 8,387). This is consistent with the main results that the intervention led to very limited behavioral changes. Users in the random sample survey were asked if they thought that wealth should be redistributed, and the surveyor explained what a wealth tax would mean, in the telephonic surveys. Respondents could say 'Yes,' 'No,' or 'Don't know.' The uncertain responses were dropped before computing these percentages, standard errors, and p-values. Based on these responses, I also created a progressiveness index, from respondents' answers to different questions relating to affirmative action and wealth redistribution. Details of the survey instrument are contained in a companion paper. Respondents were further divided into toxic and non-toxic groups, based on their exposure to toxic content at baseline in the admin data. If a user's exposure to toxic content was above the median level at baseline, they were classified as a toxic user. Each group was balanced in terms of treatment status, on account of the random assignment and sample selection.

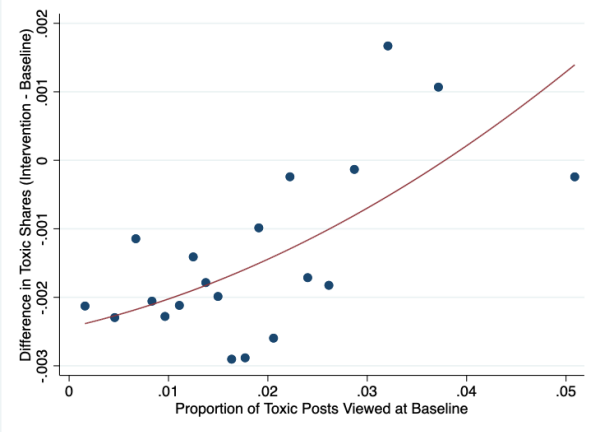
Figure D.19: Structural Estimates and Validation



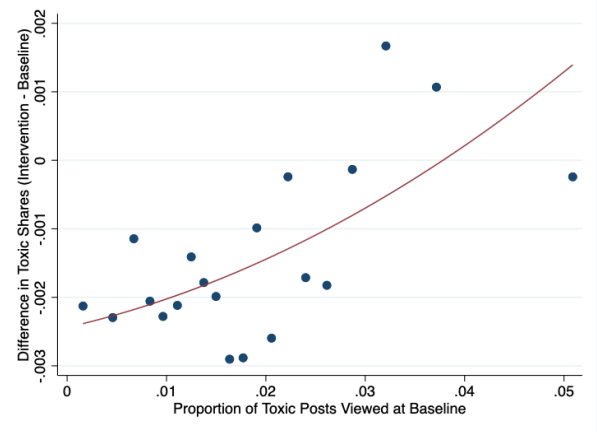
(a) Baseline views and intervention period shares in the treatment group



(b) Intervention period views and shares in the treatment group



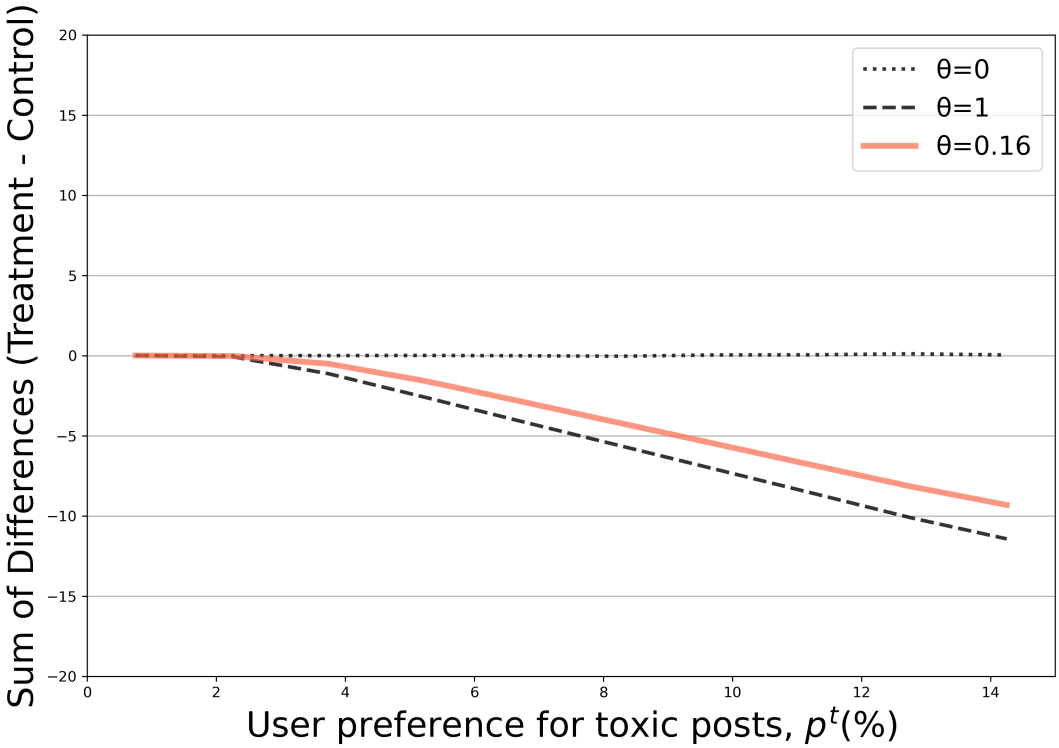
(c) Baseline views and intervention period shares in the control group



(d) Intervention period views and shares in the control group

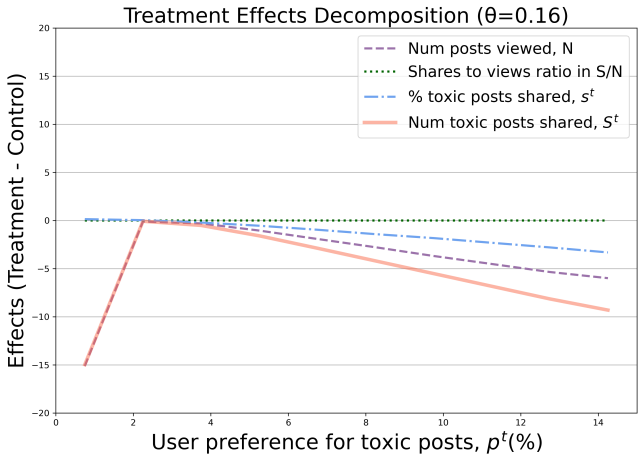
Notes: Panel (a) shows that $\gamma_1 = -\theta$ is negative, and the relationship between differences in toxic shares and toxic views at baseline approximates a linear one, as predicted by the structural model. Panel (b) shows that the relationship between differences in toxic shares (from baseline to intervention period) and in the toxic views during the intervention, produces a relationship that can be positive, as well as distinct from $-\theta$. This is because the estimation strategy uses proportion of toxic views at baseline. The intervention period variation in toxic views is concentrated around the mean, by design of the intervention. As a result, this variation is not informative about the rate at which users update their behavior according to the perceived behavior of others, or the prevalent social norms. Panels (c) and (d) reiterate that the relationship between toxic views and differences in toxic shares, in the control group, do not convey any meaningful information because control users are always in steady state. This means that the said relationship is not estimable in the control group. The binscatter plots constructed using the control group data are distinct from the main plot in panel (a).

Figure D.20: Treatment effects on total number of toxic posts shared for different influence factors, θ

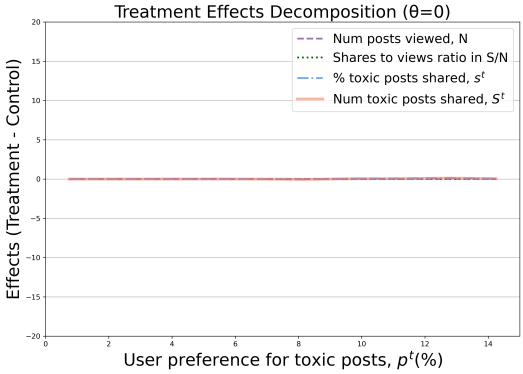


Notes: This figure shows that the simulated treatment effects on number of toxic posts shared is negative for more toxic users, when the rate at which exposure influences behavior is $\theta = 0.16$, as estimated using the structural model and the empirical distributions of various outcomes. This shows that, for the parameter values calibrated using the method of matching moments (See Appendix C.3 for details), the structural model correctly predicts that the treatment effect on the number of toxic posts shared is negative for toxic users. The treatment effect is then simulated for different influence regimes: $\theta = 0$, when users share content *mechanically*, and $\theta = 1$, when users are fulling *malleable*. The treatment effect on the number of toxic posts shared is constant at zero, in the case of mechanical users (i.e. $\theta = 0$). However, when $\theta = 1$, users with lower proclivity to toxic content share more toxic content, because they are fully influenced by the content they are exposed to. Note that, the sharp decrease in the predicted treatment effect when $\theta = 0.16$ is driven by the model prediction that changes in overall engagement with the platform are symmetric across extreme users.

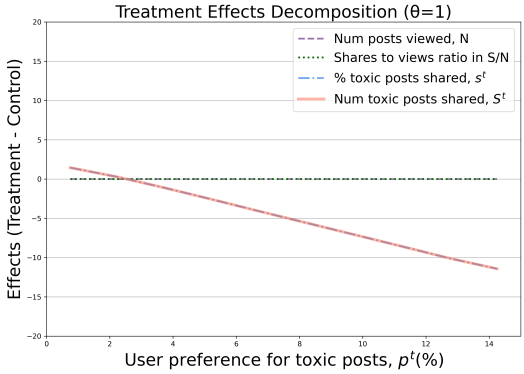
Figure D.21: Decomposition of treatment effects, in different updating regimes



(a) Estimated $\theta = 0.16$



(b) Mechanical users, $\theta = 0$



(c) Malleable users, $\theta = 1$

Notes: This Figure shows that if users were updating their behavior at the same rate θ, the decrease in the number of toxic posts shared is largely driven by the disengagement effect, especially for more toxic users (on the right extreme of the p^t distribution). It decomposes the treatment effect into its two constituent parts, namely, the engagement effect, on number of posts viewed N, as well as the shares to views ratio S/N, and the influence effect, on the probability of sharing toxic content s^t. Panel (a) shows that the reduction in total views (or the disengagement effect) has a higher contribution to the reduction in toxic shares, than the reduction in the probability of sharing toxic content. Panel (b) shows that there is no change in behavior if users were completely mechanical ($\theta = 0$). Panel (c) shows that treatment effect is entirely driven by the influence effect if users were completely malleable, and that the number of toxic posts would increase for non-toxic users if $\theta = 1$. The model generated simulated outcomes that are consistent with the data, on the right side of the p^t distribution. The behavior of non-toxic users is not predicted by the model with constant θ across users, and is consistent with the idea that non-toxic users are not as malleable.

E Supplementary Tables

Table E.1: Regression results for all outcome variables

	Num Logins	Time Spent (in hours)	Num Posts Viewed
Treatment	-1.270**	-2.531**	-35.497**
	(0.042)	(0.584)	(2.208)
Control Mean	21.594**	7.104**	246.654**
	(0.021)	(0.583)	(1.361)
	Time Spent per Post	Num Posts Shared	Shares to Views Ratio
Treatment	-0.053**	-6.367**	-0.114**
	(0.002)	(0.206)	(0.007)
Control Mean	0.127**	18.396**	0.261**
	(0.001)	(0.131)	(0.004)
	Prop Activity on Weekends	Prop Activity during Daytime	Num Searches per Post Viewed
Treatment	0.010**	-0.035**	0.016**
	(0.001)	(0.002)	(0.001)
Control Mean	0.261**	0.214**	0.104**
	(0.001)	(0.001)	(0.001)
	Prob Leaving Platform	Num Toxic Posts Viewed	Perc Toxic Posts Viewed
Treatment	0.006**	-5.024**	-0.641**
	(0.001)	(0.172)	(0.033)
Control Mean	0.030**	18.806**	7.416**
	(0.000)	(0.129)	(0.018)
	Num Toxic Posts Shared	Perc Toxic Posts Shared	Tox Share to Tox View Ratio
Treatment	-0.093**	0.120**	0.007**
	(0.010)	(0.038)	(0.001)
Control Mean	0.474**	1.547**	0.040**
	(0.006)	(0.018)	(0.001)
N		231814	

Notes: This table shows that the intervention caused disengagement with the platform, by showing negative and significant estimates of treatment effects on total number of posts viewed and shared, number of times users logged on, and total time spent. Each cell estimates the following regression equation with different outcomes (Y_i), $Y_i = \beta_0 + \beta_1 D_i + \varepsilon_i$. The average user viewed and shared fewer toxic posts, but the proportion of toxic posts shared increased. This table also shows that the intervention increased users' search costs of using the platform, as measured by the number of searches performed. This could explain why the treatment effect on proportion toxic shares is positive, despite the treatment effect on proportion toxic views being negative. Robust standard errors in parenthesis. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table E.2: Regression results for different thresholds criterion for toxic content

Average Views				
	(1)	(2)	(3)	(4)
	Continuous Score	0.2 Threshold	0.3 Threshold	0.4 Threshold
Treatment	-0.824***	-0.493***	-0.00881	-0.171***
	(0.013)	(0.027)	(0.017)	(0.010)
Control Mean	5.930***	4.434***	2.004***	0.697***
	(0.006)	(0.013)	(0.009)	(0.005)
Average Shares				
	(1)	(2)	(3)	(4)
	Continuous Score	0.2 Threshold	0.3 Threshold	0.4 Threshold
Treatment	0.112***	0.025	0.0333	0.052***
	(0.015)	(0.026)	(0.018)	(0.013)
Control Mean	1.289***	0.839***	0.405***	0.227***
	(0.008)	(0.013)	(0.009)	(0.007)
N			231814	

Notes: This Table shows that the treatment effect on the proportion of toxic posts viewed and shared, when toxicity is defined using the continuous toxicity score (1), and different thresholds for the binary toxicity score (2-4). All the results are consistent with the main results on toxic exposure in Table E.1. Standard errors are robust at user level. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table E.3: User characteristics correlated with the probability of leaving the platform

Variable	Coefficient	Interaction Coefficient
Treatment Effect	0.014	N/A
	(0.008096)	N/A
Number of Views (Baseline)	-0.000***	0.000
	(0.000)	(0.000)
Number of Shares (Baseline)	-0.000	-0.000
	(0.000)	(0.000)
Toxic Shares (Baseline)	0.000*	-0.000
	(0.000)	(0.000)
Toxic Views (Baseline)	-0.000***	0.000
	(0.000)	(0.000)
Male Gender	-0.004***	-0.002
	(0.001)	(0.002)
Days since account created	0.000***	-0.000
	(0.000)	(0.000)
User Age	-0.000	-0.000
	(0.000)	(0.000)
Proportion content viewed on weekends	-0.002	0.003
	(0.002)	(0.004)
Proportion content shared during daytime	0.002	-0.001
	(0.001)	(0.003)
Share of views in Bollywood Genre	0.039***	0.008
	(0.005)	(0.011)
Share of views in Devotion Genre	0.015***	0.006
	(0.004)	(0.009)
No Assigned Genre	0.049***	0.037
	(0.011)	(0.023)
Share of views in Greetings Genre	0.028***	0.001
	(0.004)	(0.008)
Share of views in Humor Genre	0.054***	-0.018
	(0.008)	(0.015)
Share of views in News Genre	0.011	-0.016
	(0.007)	(0.013)
Share of views in Politics Genre	-0.014	-0.011
	(0.032)	(0.070)
Share of views in Romance Genre	0.054***	-0.002
	(0.005)	(0.010)

Notes: This Table shows that, conditional on observable user characteristics, treatment assignment is not correlated with the probability of leaving the platform. This also shows that the treatment does not differentially impact the probability of leaving the platform, for given observable user characteristics. This means that the treated leavers are not systematically different from the control leavers. These results are obtained by estimating the regression equation $\mathbb{1}_i(\text{left platform} = \text{ yes}) = \beta_0 + \beta_1 D_i + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \sum_c \beta_{1c} D_i \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where $\mathbb{1}_i(\text{left platform} = \text{ yes})$ is an indicator taking value 1, when user i leaves the platform. Column (1) reports estimated β_c's, while column (2) reports estimated β_{1c}'s. Standard errors are robust at user level. $p < 0.05^*, p < 0.01^{**}, p < 0.001^{***}$.

Table E.4: Structural estimates using OLS regressions

	(1) Proportion Toxic Posts Shared (Intervention - Baseline)
Proportion Toxic Posts Viewed	-0.104**
	(0.037)
N	63041

Notes: This table shows that the structural estimates of θ obtained using an OLS regressions are biased downwards. Dependent variable is differences in differences between probability of sharing toxic and non-toxic content, between intervention period and baseline, for treated users only. The explanatory variables are constructed by averaging differences between proportion of toxic and non-toxic posts viewed by treated users. Robust standard errors in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

F Robustness to Attrition

Table F.1 reports the estimated treatment effects of the intervention on various outcome variables. Throughout the paper I have maintained that the relevant value for users who stop coming to the platform, or leave it entirely, is zero. This is true for outcomes including the number of posts shared/ viewed, the number of toxic posts shared/ viewed, and the proportion of toxic posts shared/ viewed. Similarly, the time spent on the platform is zero for users who leave the platform.

Table F.1: Heterogeneous Treatment Effects

Quantile	Number of Logins		Number of Shares		Number of Views	
	Control Mean	Effect (SE)	Control Mean	Effect (SE)	Control Mean	Effect (SE)
Q1	20.929	-1.225 (0.114)	22.838	-4.752 (0.688)	226.650	1.063 (5.82)
Q2	21.820	-1.472 (0.113)	24.629	-8.165 (0.685)	272.036	-15.298 (5.819)
Q3	22.101	-1.521 (0.112)	22.386	-7.435 (0.608)	304.601	-52.926 (6.521)
Q4	22.152	-1.483 (0.111)	19.508	-6.994 (0.522)	325.085	-45.855 (7.785)
Q5	22.376	-1.281 (0.111)	14.311	-5.456 (0.393)	328.739	-76.34 (6.155)

Quantile	Time Spent (in hours)		Num of Toxic Shares		% Toxic Shares	
	Control Mean	Effect (SE)	Control Mean	Effect (SE)	Control Mean	Effect (SE)
Q1	4.647	-1.151 (0.089)	0.329	0.009 (0.017)	0.992	0.255 (0.081)
Q2	6.441	-1.972 (0.564)	0.462	-0.042 (0.022)	1.248	0.108 (0.081)
Q3	9.206	-4.264 (2.592)	0.584	-0.092 (0.027)	1.533	0.168 (0.092)
Q4	9.759	-4.446 (2.613)	0.695	-0.186 (0.032)	1.951	0.058 (0.104)
Q5	7.360	-2.051 (0.137)	0.722	-0.246 (0.034)	2.707	-0.278 (0.126)

Quantile	Ratio of Toxic Share to View		Num of Toxic Views		% Toxic Views	
	Control Mean	Effect (SE)	Control Mean	Effect (SE)	Control Mean	Effect (SE)
Q1	0.043	0.003 (0.003)	9.579	3.352 (0.299)	4.998	1.105 (0.081)
Q2	0.042	0.005 (0.004)	15.038	0.729 (0.36)	6.000	0.38 (0.078)
Q3	0.041	0.009 (0.004)	21.235	-5.011 (0.483)	7.043	-0.312 (0.083)
Q4	0.039	0.004 (0.003)	28.047	-9.263 (0.608)	8.319	-1.349 (0.078)
Q5	0.034	0.016 (0.005)	37.581	-18.573 (0.641)	10.775	-2.989 (0.087)

Notes: The table reports the estimated treatment effects of the intervention on the outcome variable, by the amount of toxicity user was exposed to at baseline, which is a proxy for their type. The treatment effect is estimated using a linear regression model, with the outcome variable as the dependent variable, and the treatment indicator as the independent variable, both aggregated at the user level. The treatment indicator is a dummy variable that takes the value of 1 if the user is treated, and 0 otherwise. The table also reports the standard errors of the estimated treatment effects. The standard errors are robust.

This may raise concerns that selective attrition could bias the estimated treatment effects, if the treated users who leave the platform are systematically different from those who stay. To test that treated leavers are not systematically different from treated stayers, I first estimated the treatment effect on the probability of leaving the platform. Although I find differential attrition by treatment status, controlling for various observable characteristics

corrects for this bias. This means that upon controlling for user attributes that are correlated with the probability of leaving among the treated, there is no selection in the probability of leaving among treated users. This is seen in Table E.3.

Table F.2: Lee Bounds for Estimated Treatment Effects

Outcome	Quantile	Treatment Effect	Standard Error	Lower Bound	Upper Bound
Num of Posts Viewed	Q1	1.063	5.820	-1092.278	2.627
	Q2	-15.298	5.819	-1412.702	-14.473
	Q3	-52.926	6.521	-1729.319	-54.646
	Q4	-45.855	7.785	-1844.840	-46.813
	Q5	-76.340	6.155	-1904.051	-79.718
Num of Toxic Posts Viewed	Q1	3.352	0.299	-45.471	3.686
	Q2	0.729	0.360	-81.304	0.898
	Q3	-5.011	0.483	-132.070	-5.222
	Q4	-9.263	0.608	-178.453	-9.727
	Q5	-18.573	0.641	-247.830	-19.658
Num of Posts Shared	Q1	-4.752	0.688	-163.773	-4.978
	Q2	-8.165	0.685	-190.771	-8.591
	Q3	-7.435	0.608	-176.153	-7.815
	Q4	-6.994	0.522	-158.755	-7.358
	Q5	-5.456	0.393	-123.217	-5.753
% Toxic Posts Viewed	Q1	1.105	0.081	-15.512	1.228
	Q2	0.380	0.078	-15.785	0.454
	Q3	-0.312	0.083	-16.999	-0.283
	Q4	-1.349	0.078	-17.639	-1.384
	Q5	-2.989	0.087	-21.795	-3.132
Num of Toxic Posts Shared	Q1	0.009	0.017	-3.398	0.012
	Q2	-0.042	0.022	-4.999	-0.042
	Q3	-0.092	0.027	-6.509	-0.095
	Q4	-0.186	0.032	-7.965	-0.194
	Q5	-0.246	0.034	-8.568	-0.259
% Toxic Posts Shared	Q1	0.255	0.081	-11.275	0.283
	Q2	0.108	0.081	-14.687	0.126
	Q3	0.168	0.092	-17.297	0.191
	Q4	0.058	0.104	-21.491	0.078
	Q5	-0.278	0.126	-30.113	-0.280
Ratio of Toxic Shares to Views	Q1	0.003	0.003	-0.497	0.004
	Q2	0.005	0.004	-0.494	0.006
	Q3	0.009	0.004	-0.489	0.010
	Q4	0.004	0.003	-0.457	0.005
	Q5	0.016	0.005	-0.398	0.017

Notes: The table reports the Lee bounds for the estimated treatment effects of the intervention on the main outcome variables. The Lee bounds are constructed using the rate of attrition, which is computed using the inverse probability of logging on to the platform. The table shows that the Lee bounds for the treatment effects are tightly estimated.

However, there still may be concerns that the estimated treatment effects are biased due to selective attrition, if the treated users who leave the platform are systematically different from those who stay, on unobservable characteristics. To address this concern, I construct Lee bounds for the estimated treatment effects, with respect to all the outcome variables (Lee, 2009). The rate of attrition is computed using the inverse probability of logging on to

the platform, and is used to construct the bounds. Table F.2 shows that the Lee bounds for negative treatment effects are tightly estimated.

G Contextual Details

This Appendix provides the contextual background that makes this study highly timely and relevant. The context of this study is India, which is the second-largest market for social media platforms. However, the implications of the study are global, as the problems of misinformation and hate speech are universal (Avalle et al., 2024).

G.1 Social Media and Indian Politics

As more and more Indians get connected to the Internet, they are more likely to be exposed to misinformation in an already polarized society. As a result, social media has been linked to organized hate crimes against minorities in India. The 2015 mob lynching of Mohammad Akhlaq, a Muslim farm worker, just outside of the National Capital Region of Delhi, highlighted the role that platforms like WhatsApp play in spreading misinformation and exacerbating hate (Arun, 2019). This unfortunate incident is by no means an isolated one, making it especially important to study the factors that drive online political divisions in India.

Social media platforms like WhatsApp, Facebook, and Twitter face an unprecedented challenge of moderating content in this massive market. Attempts at moderating social media in the US have met with loud criticism from both sides of the political spectrum (Kominers and Shapiro, 2024). This task is even more difficult in India because these are American companies operating in a vastly different context, where hate speech on social media propagates in very atypical ways. The enormity of this task was most recently highlighted by Meta's inability to control anti-Muslim disinformation campaigns, just ahead of the Indian election of 2024.[29]

Context-driven content moderation is a difficult challenge, also because the production of hate in the Indian context is very often linked with institutions that enable these platforms to do businesses. This was seen, for instance, when Twitter suspended various accounts linked with the Farmer's Movement during massive protests against the controversial farm bills passed by the Indian Parliament (Dash et al., 2022). Similarly, the Wall Street Journal has alleged that Facebook India's Public Policy Head selectively shielded offensive posts of leaders of the ruling Bharatiya Janata Party (BJP), which has been variously described as Prime Minister Modi's Hindu Nationalist Party[30].

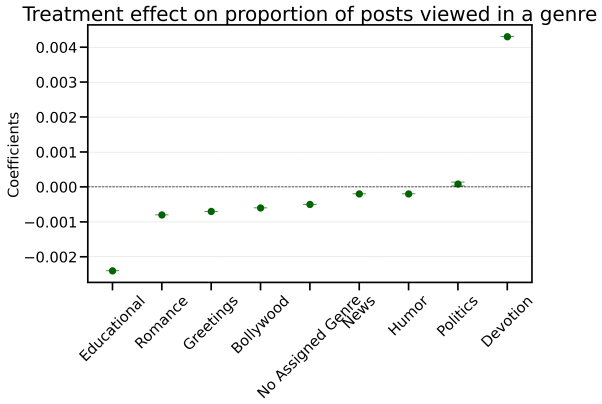
[29]See `https://thewire.in/tech/meta-approved-ai-manipulated-political-ads-during-indias-election-report`

[30]WSJ has alleged that BJP leader, T. Raja Singh, has said in Facebook posts that Rohingya Muslim immigrants should be shot, called Muslims traitors and threatened to raze mosques to the ground. PM Modi's BJP has, in many instances, encourages blatant calls for violence against the country's largest religious majority, i.e. Muslims.

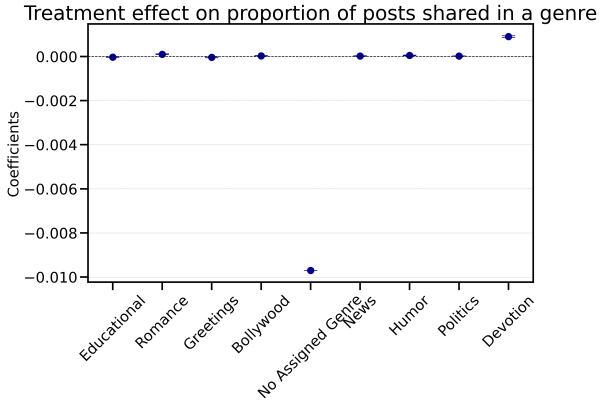
G.2 SM: 'Indian TikTok'

SM is one of the most popular platforms in the country, as users can create and share content in over a dozen regional languages. On this platform, users interact with content generated by other users, who are typically super-stars or influencers in a particular genre, on the platform. Super star content creators could be comedians, dancers, or singers, who are some times supported by the platform, to enhance engagement.[31] While the platform is home to organic content creators, various politicians, and Bollywood celebrities also sometimes interact with their follower base on this platform.

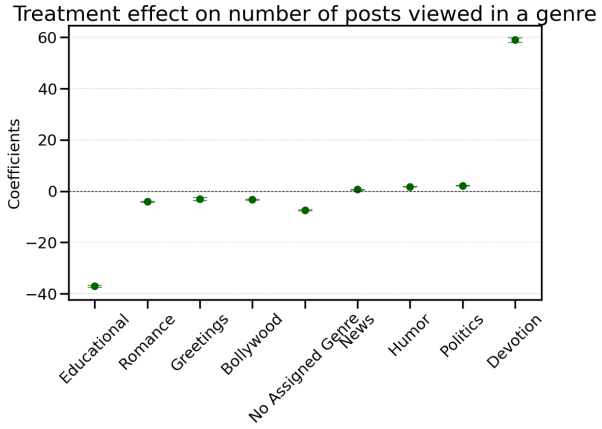
Figure G.1: Treatment effects on viewing and sharing content from various genres



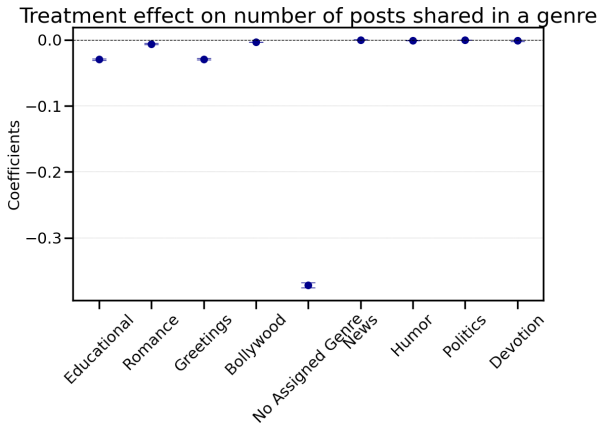
(a) F-stat: 6122.932, p-value: 0.000



(b) F-stat: 1317.659, p-value: 0.000



(c) F-stat: 451.112, p-value: 0.000



(d) F-stat: 276.637, p-value: 0.000

Notes: These plots show that the treatment affected the number of posts shared and viewed in different genres. Although there was a large increase in exposure to devotional or religious content, the treatment effect on number and proportion of religious posts shared was much smaller. The treatment effect on views in educational, romance, bollywood, and greetings genres was negative. However, there was no commensurate decrease in the number of posts shared in these genres. Standard errors are robust at user lebel, and are computed at the 5% level of significance.

[31] See, for instance, the Instagram profile of 'India's First Trending Transgender Model', who rose to fame through her dance videos on SM: `https://www.instagram.com/khushi1216/?hl=en`.

Content based social networks, such as SM, are centered around topics like Politics, Religion, and Good Morning (or Greetings) messages. Religious posts (both relating to Islam and Hinduism) are by far the most popular genre on the platform. India's young population seem to seek out relationship and dating advice, while older populations seem more invested in motivational content. Figure G.1 provides details on the treatment effects on the popularity of various genres on the platform.

Politics is the least favored genre on the platform, but 20% of the content in this genre was classified as toxic, during the first month of the intervention. I used the Perspective API to classify content as toxic or non-toxic, irrespective of the genre it belonged to. Posts are automatically classified into broad genres in the data, potentially using the user generated hash-tags associated with each post. The algorithms used to classify content were not disclosed by the platform to this author.

The interactions on SM are mostly conducted through the 'trending' feed, which is also the landing page when a user logs onto the platform (See Figure G.2). In this way, the platform's interface resembles that of TikTok, than the more widely studied platforms like X (formerly, Twitter). User interaction in this network is possible only because of the similarity in content that users have shown to engage with. Therefore, SM is distinct from platforms like Facebook, where users engage with content from 'Friends' or from 'Groups.'

Figure G.2: Landing page and trending tab on SM

Notes: This image shows the landing page and trending tab on the social media platform, SM. Users see a feed of image posts and the creator generated hashtags on the landing page, much like Instgram. Users can share, comment, like, or download the post to their phones. Sharing refers to sharing on WhatsApp, and not on the platform itself, for instance on user's own profile. This makes SM's interface very different from other platforms like X (formerly, Twitter), where users can share posts with their followers, through their profile on the platform. A user can see other users who liked and commented on a post, but not the users who shared the post. SM posts are classified into broad categories or genres like 'politics' (in this image), 'devotional,' 'romance,' 'Bollywood,' 'greetings,' and 'educational.'

H Matrix Factorization Model

Matrix Factorization algorithms provide some approximation of user preferences from their previous engagement with posts on the platform. This is done with the objective of optimizing user retention and engagement by serving them the type of content they have shown affinity towards in the past. The algorithm factorizes a matrix of engagement at the user-post level for some abstract set of user and post features.

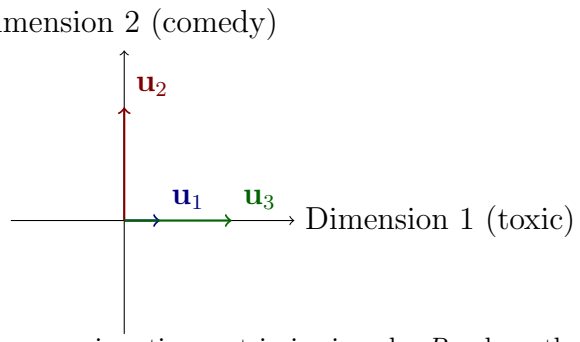
Figure H.1: An example of SVD decomposition into two-dimensional user embeddings U, eigenvalues Σ, and movie embeddings V^T

(a) Matrix representation of vector embeddings

$$R = \begin{pmatrix} 1 & 0 \\ 0 & 2 \\ 3 & 0 \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{10}} & 0 \\ 0 & 1 \\ \frac{3}{\sqrt{10}} & 0 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$U \qquad \Sigma \qquad V^T$$

(b) Graphical representation of vector embeddings



Notes: In this example, a user-movie rating matrix is given by R, where three users rate two movies on a scale of 1 to 5. The idea is to learn user tastes in some low-dimensional space of latent features. This is because the dimensionality of the R matrix rises with the number of users and movies. Singular Value Decomposition (SVD) breaks this matrix down as **(1)** U represents the user embeddings (u1 and u2), showing how users relate to the abstract features; **(2)** Σ is a diagonal matrix containing singular values (σ_1 and σ_2), which scale the importance of each feature; **(3)** V^T represents the movie embeddings (v1 and v2), showing how movies relate to the abstract features. By multiplying U, Σ, and V^T back together, the original matrix R is reconstructed. The embeddings in U and V are plotted in a 2D space to visualize their relationships. These plots show that the first user is more interested in the first movie (or movies of that type), while the second user is more interested in the second movie (or movies of that type). The two dimensions represent abstract features that summarize the original data's structure and relationships. For example, dimension 1 could represent the toxic genre, while dimension 2 could represent the comedy genre.

H.1 Illustration: Control Algorithm

Consider an example with three users and two movies in Figure H.1. I use singular value decomposition (SVD) to factorize the engagement matrix into two-dimensional user and post

latent features. If we interpret dimension 1 of the factor matrices as movies relating to toxic genre, and dimension 2 as movies relating to comedy genre, then the factorization process generates a vector of weights for each user with respect to these attributes. In this example, the weights (or embeddings) reveal that users 1 and 3 have a higher proclivity for toxic movies, while user 2 is likely to rate comedy movies higher. As a result, these attribute weights enable a platform to serve toxic movies to users 1 and 3, and comedy movies to user 2, in order to maximize user satisfaction.

More generally, this factorization process generates a vector of weights for each user with respect to some post attributes, so that a cross product of weights for user and post latent features gives the predicted engagement matrix, or the scores that generate ranking of various posts that each user is recommended in the future. These vector-weights in the space of some latent post/ user features are known as embeddings in the machine learning literature (Athey et al., 2021). The user features produced are latent representations of user behavior revealed in the past, and are produced by minimizing a known loss function using Stochastic Gradient Descent. These latent features are represented as a multi-dimensional embedding vector, where each element in the vector represents the weight each user is predicted to put on some latent post attributes.

H.2 Illustration: Treatment Algorithm

In this experiment, the content recommendations for the control group are generated as per the usual personalization algorithm. For the treatment group, the algorithm is modified to replace actual user embeddings with randomly selected user embeddings from the control group distribution. In the example below, user 2 is randomly chosen to be treated, and the embeddings for user 2 are replaced with the average of the embeddings for users 1 and 3.

Figure H.2: Matrix representation of vector embeddings, for treated and control users

$$\begin{pmatrix} \frac{1}{\sqrt{10}} & 0 \\ \rho_{21} & \rho_{22} \\ \frac{3}{\sqrt{10}} & 0 \end{pmatrix} \cdot \begin{pmatrix} \sqrt{10} & 0 \\ 0 & 2 \end{pmatrix} \cdot \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}$$

$$U \qquad\qquad \Sigma \qquad\qquad V^T$$

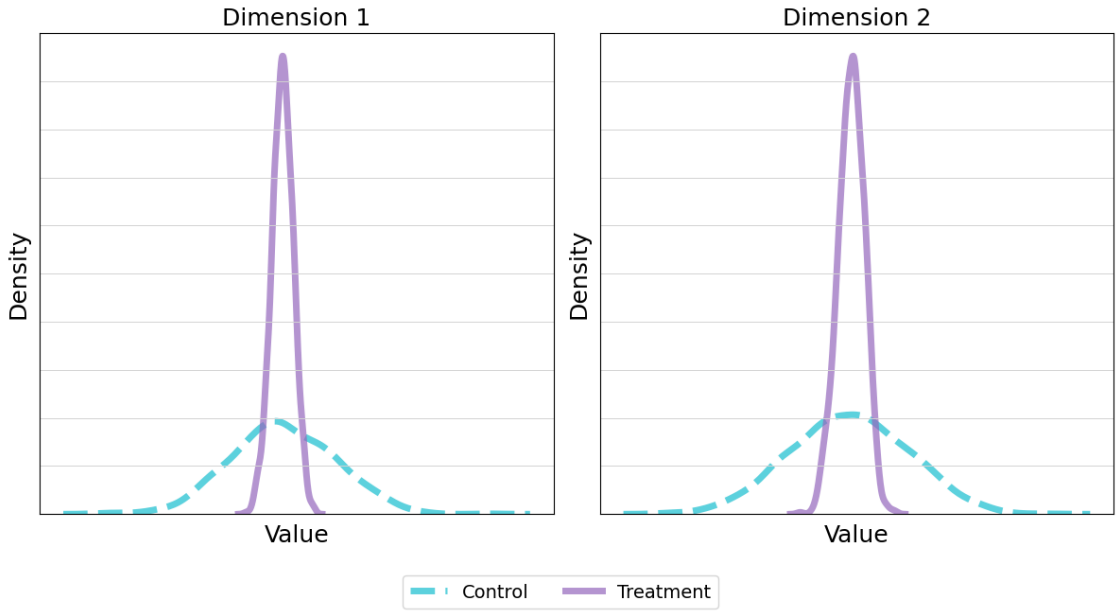
Notes: This figure shows the user embeddings for the control group (in black) and the treatment group (in red). The treatment group embeddings, e.g. user 2, are generated by randomly selecting from the distribution of control group embeddings. This determines the order of different types of posts that are recommended to each user.

The embeddings generated for each treated user are equal to the average of the embeddings for the control group users. Therefore, there is not enough variation in the embedding assignment within the treatment group, as the treatment embeddings are concentrated on the mean embedding value, by application of the Law of Large Numbers (LLN). This is depicted in Figure H.3, for the simulated (two-dimensional) recommendation algorithm. This necessitates the need for a structural model to identify the effect of exposure on engagement.

It may be expected that in bringing the treatment group embeddings closer to the mean, the treatment biases content exposure among the treated towards more popular posts. This

is because the average user's embeddings are likely to be closer to the preferences of the largest number of users on the platform, making them more popular. However, Table H.1 shows that the treatment group was exposed to less popular posts than the control groups because the random numbers picked to generate preference weights for the treatment group were not representative of any actual user preferences on the platform.

Figure H.3: Distribution of simulated two-dimensional embedding vectors



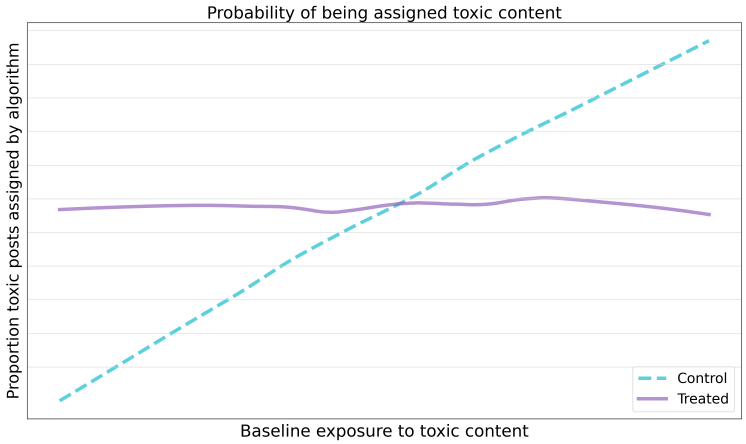
Notes: This graph shows that the two dimensions (components) of the embedding vector follow a Gaussian distribution, where the embeddings were simulated using a simple SVD algorithm and a matrix of engagement in the control group. An embedding is a representation of complex data in a lower-dimensional space. The dimensions of these vectors are abstract features that summarize the original data's structure and relationships. Then, the randomly selected embeddings for the treated users are centered around the mean of each embedding dimension, and the spread of control user embeddings is larger than the embeddings generated for treated users. This is because the treatment embeddings are drawn uniformly at random, each day, from a given sample of control embeddings during the intervention period (LLN).

Table H.1: Popularity of posts viewed by users in the treatment group

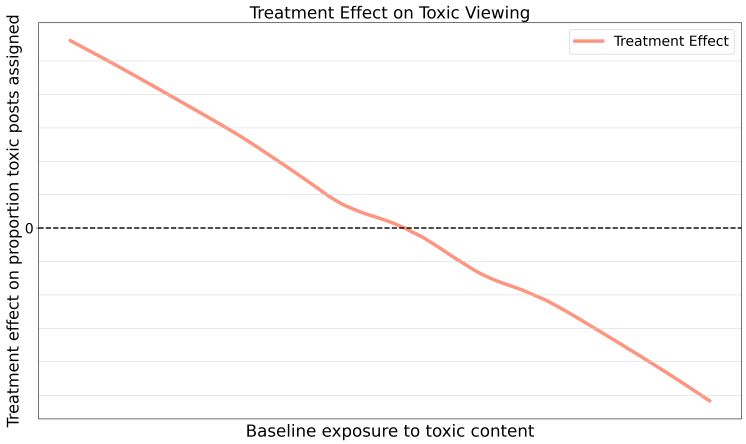
	Views on posts viewed	Likes on posts viewed	Shares on posts viewed
Treatment	-140732.408**	-1549.188**	-3966.425**
	(758.901)	(7.527)	(28.271)
Constant	241586.576**	3093.363**	5999.583**
	(682.964)	(6.491)	(26.112)
Obs		231814	

Notes: This Table shows that, contrary to expectations, the treatment group was exposed to less popular posts than the control group. It is possible that in bringing the treatment group's preference weights closer to those of an average user, the intervention recommended posts are more appealing to the widest audience. However, this is not observed in the data. Standard errors are robust at user level. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Figure H.4: Example of correlation between simulated user preferences and recommendations from a simulated personalization algorithm



(a) Distribution of the first dimension of the embedding vector across treatment and control



(b) Treatment effect on the embedding values assigned, sorted by baseline embedding value

Notes: This Figure shows that there is a positive correlation between the user preferences (measured using embedding vectors at baseline), with the type of posts recommended by a simple personalization algorithm. The algorithm used to simulate the embeddings for both treatment and control groups uses Singular Value Decomposition to factorize a simulated matrix of engagement. This generates two-dimensional embedding vectors for each user and each post, where each dimension users' preference weights on different post attributes, e.g. tragedy, toxicity, comedy, etc. To fix ideas, this graph shows the first dimension of the embedding vector, which represents the toxicity of the post (as an example). In breaking this correlation between user preferences and the preferred content, treatment is expected to have a smaller effect (in absolute terms) on users with embeddings closer to the average, at baseline. This is because the treatment algorithm assigns toxic content with the average probability in the control group, as the treated users are simply assigned the average control embedding (as shown by the flat curve in panel (a)). On the other hand, users with more extreme preferences had bigger absolute effects in content exposure. Embeddings from the treatment group were uniformly drawn from an epsilon ball centered around the mean control embedding. Therefore, the embedding values for the control users form an upward sloping curve, with respect to user preferences for toxic content (which is the first dimension of the embedding vector). There is no correlation between the user embeddings in the treated groups, and users baseline embeddings, by design of the experiment.

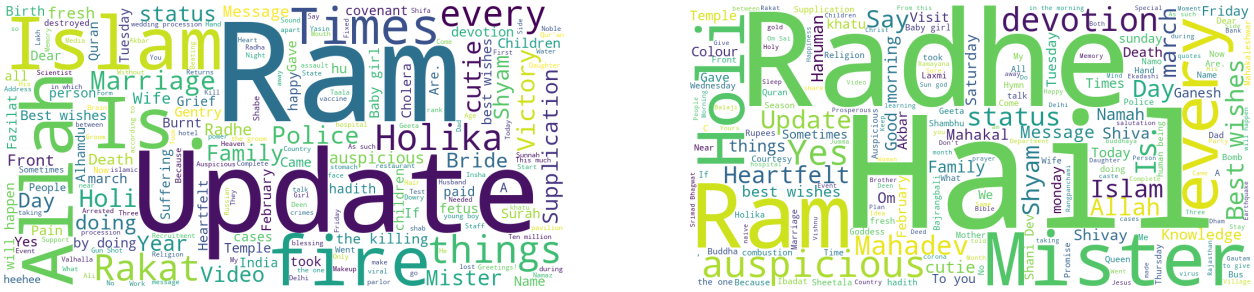
I Text Analysis

The post data is characterized by broad tag genres, employing user generated hashtags. The administrative data also consists of text on the images/ videos in the user generated posts, that was obtained through an automated optimal character reader (OCR). This is a rich source of information, and I adopt various methods to analyze the text data, in order to understand the qualitative nature, tone, and political slant of these posts.

I.1 Tokenization, Word Clouds, and Topic Models

I begin describing the text data by translating from the original Hindi, and summarizing the most common words in the political posts in Figure I.1. This summary measure is based on more than 20 million posts that were viewed and shared by users in the baseline and intervention periods. The text analysis currently excludes a dozen other Indian regional languages in which users can consume content.

Figure I.1: Word clouds depicting words associated with highly toxic posts



(a) High Toxicity (b) Low Toxicity

Notes: This Figure shows word clouds constructed using the TF-IDF vectorizer, on posts classified into high and low toxicity categories respectively. Cut-off to classify posts into high and low toxicity categories is 0.2, based on the toxicity scores provided by Perspective API. The figure demonstrates overlap in words pertaining to religion in both categories, for example 'Islam' and the Hindu mythological god-king 'Ram,' who is also central to Hindu nation building agenda of the current ruling government. This highlights the need for contextual embeddings to characterize the text data. Perspective's toxicity algorithm uses human labelled comments and BERT models to provide toxicity scores to each post, by representing posts in some latent space as embedding vectors.

Figure I.1 shows that the most common word in posts labelled as toxic is 'Ram,' which is a reference to legendary Hindu deity, who is said to have blessed the Hindu Nationalist project.[32] The Hindu nationalist project is a political ideology that is associated with the ruling party in India, that has been accused of promoting anti-minority sentiments, and even promoted outright calls for ethnic cleansing in extreme instances (Jaffrelot, 2021).

[32]For instance, see Kalra (2021) for details on a coordinated campaign carried out in the name of Lord Ram, that was aimed at inciting violence against Muslims in different parts of India. This campaign, the *Ram Rath Yatra*, was a precursor to the 1992 Babri Masjid demolition. The temple built in place of this mosque was inaugurated by the current Prime Minister of India, Narendra Modi, in Junuary 2021. See `https://www.bbc.com/news/world-asia-india-68003095`

However, I find a significant overlap in the most common words across posts that were classified as toxic or not. For instance, the words 'Ram,' 'Islam,' 'Allah' are common in both toxic and non-toxic posts. This demonstrates that analyzing tokenized vector of words may lead to misleading conclusion. The text analysis must include sufficient information about the context in which the words are used. Therefore, I tried to gather a better sense of the context in which the words were used, by employing topic models on the text data.

The LDA and BERT topic models provide useful information about the context in which the words are used, but the variation in topics, especially in the Politics genre, was too limited to be useful. Since, I am interested in the harm that posts can cause, I currently limit my analysis to hatefulness or toxicity of posts. This is a task best suited for some off-the-shelf classification algorithms, that I describe later. Therefore, I use semi-supervised Machine Learning methods that take contextual embeddings into account, while achieving a narrower objective: classifying posts as toxic or not.

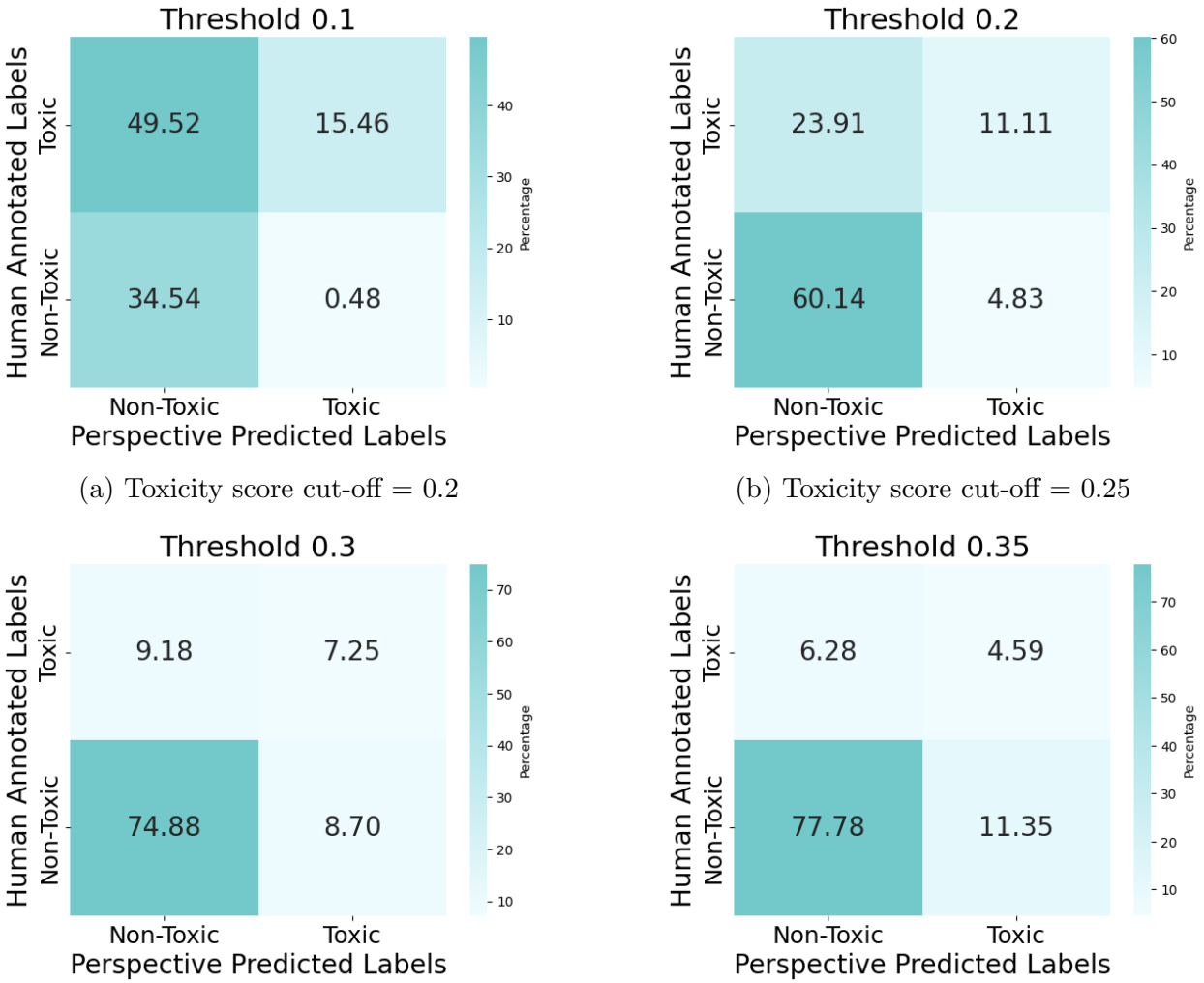
I.2 Toxicity Algorithm

In keeping with the literature on social media harms, I use the Perspective API to classify posts as toxic or not (Jiménez Durán et al., 2023). The Perspective API is a machine learning algorithm developed by Jigsaw at Google, that provides a machine learning solution to detect posts that are likely to harm a participant in a discussion. I provide examples to illustrate the toxicity classification algorithm in Table I.1.

Figure I.1 shows the most commonly occurring words (in English) across posts that were classified as toxic or not, and the overlap in words across the two groups. The overlap in words across the two groups also testifies that the toxicity scores are sensitive to contextual embeddings, that the Perspective algorithm extracts from the text data. This validates the need for contextual embeddings for text classification.

I validate the performance of this method for multi-lingual abusive speech detection by comparing results with a choice of hate speech classification algorithms and with manually annotated posts that were viewed on SM for different toxicity thresholds. The confusion matrices in Figure I.2 show that the 0.2 cut-off has the best performance in terms of correctly classifying toxic posts. This precision criterion is important because toxicity is a rare outcome and can, therefore, make automatic detection difficult (Banerjee et al., 2023). I also find that the F1 score (combining precision and recall) for the 0.2 cut-off is the highest among the four thresholds.

Figure I.2: Confusion matrices for different cut-offs in toxicity scores



(a) Toxicity score cut-off = 0.2

(b) Toxicity score cut-off = 0.25

(c) Toxicity score cut-off = 0.3

(d) Toxicity score cut-off = 0.35

Notes: These confusion matrices show Type I and Type II errors for four thresholds for classifying a post as toxic, namely 0.2, 0.25, 0.3, 0.35. User posts were assigned continuous toxicity scores using the Perspective API, and then classified as being toxic or not for the two thresholds. These scores were compared with posts annotated as hateful by two human annotators hired at Brown University. The threshold of 0.2 was chosen because toxic posts are correctly identifies at this threshold with high accuracy. I argue that this is the most important criterion for the classification task, because toxic posts are a rare occurrence in the data. I also validate this cut-off using the F1 score, which is highest for the 0.2 threshold.

Table I.1: Examples of text data (English translations) with toxicity scores

Text	**Toxicity Score**	**Toxicity Classification**
Break those rocks Jai Shri Ram which are standing in the path of religion and shoot those criminals who have dirty intentions on the women of our country	0.399	Toxic
LIVE LATEST UPDATES 0.01% population wants 'Khalistan.' 18% want 'Ghazwa-e-Hind' and 80% want cheap onions and tomatoes. It is bitter but true.	0.327	Toxic
People travelling on "Bharat Jodo" route are now facing problem with the name "Bharat" instead of India.	0.172	Non-Toxic
Mohammed Shamim's disgusting act ! Lakhs of pilgrims kept trusting Mohammed Shamim... Mohammed Shamim used to make tea from urine water and sell it. Mohammed Shamim used to run a shop in Kerala's Sabarimala temple premises.	0.479	Toxic
00 Death does not occur only when the soul leaves the body. He is also dead who remains silent even after seeing his religion and culture being attacked. 00	0.174	Non-Toxic
Giqa Bihar wire procession of thieves (temple thief) (coal thief) (fodder four) (land thief)	0.361	Toxic
Bhajanlal Sharma will be the new Chief Minister of Rajasthan.	0.008	Non-Toxic
Don't make us jokers, when Christians being 2% do not celebrate Ramnavami, why do we Hindus being 80% celebrate Christmas, joke our children on 25th December, Jai Satya Sanatan	0.361	Toxic
In this I.N.D.I.A alliance Everyone is against "Ram" and those who are not with Ram are of no use to us Jai Shri Ram	0.267	Toxic
Why has it been proved that sycophants are the biggest problem? Who is the master of sycophants? He is the biggest problem.	0.061	Non-Toxic

Notes: The table shows examples of text data in English, with toxicity scores provided by the Perspective API. The toxicity score is a continuous measure that ranges from 0 to 1, with 0 indicating healthy contributions and 1 indicating very toxic content. The Perspective API uses a mix of supervised and semi-supervised machine learning methods, and is sensitive to context while assigning toxicity scores. The Perspective API is widely used in academic research and by publishers to identify and filter out abusive comments.