Free to Choose or Programmed to Follow? Evidence from a Large-Scale RCT on Recommendation Algorithms

Aarushi Kalra*

October 6, 2024

(please click here for the most recent version)

Abstract

To what extent is engagement with radicalizing social media content driven by algorithmically curated feeds and by user tastes? I conduct an RCT replacing personalization algorithms with random content delivery for over one million users of a prominent TikTok-like platform in India. I find a trade-off between the societal benefits of the intervention and its potential impact on producer and consumer surplus: random post recommendation lowers exposure to anti-minority ('toxic') content by 27% on average, but also lowers overall platform usage by 35%, imposing substantial costs on the platform. Strikingly, the benefits were also blunted as the proportion of toxic posts shared per toxic post viewed increased by 18%, even though the aggregate number of toxic posts shared reduced by 20%. This effect was driven by users with higher proclivity to toxic content at baseline, as they sought out posts that the algorithm did not recommend to them. I rationalize these results with a model of an engagement maximizing algorithm that faces users who choose what social media to consume according to heterogeneous preferences. Estimated behavioral parameters reveal that user behavior is relatively immalleable with an elasticity of 0.16. This results in limited effectiveness of regulatory measures that target algorithms.

Keywords: Digital Economies, AI and Technology Policy, Development, Machine Learning

*Department of Economics, Brown University, Providence, RI, 02906 (email: aarushi_kalra@brown.edu).
I am grateful to Andrew Foster, Brian Knight, Stelios Michalopoulos, Peter Hull, and Daniel Björkegren for their continued guidance and support. This project has greatly benefitted from helpful discussions with Pedro Dal Bo, Bryce Steinberg, Jesse Bruhn, Matthew Peccenco, Ro'ee Levy and seminar participants at Brown University. I thank Aryan Srivastava, 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 National Science Foundation Dissertation Research Improvement Grant, Weiss Family Fund for Research in Development Economics, as well as Orlando Bravo Center for Economic Research and Saxena Center for Contemporary South Asia, both housed at Brown University. All remaining errors are my own.

1 Introduction

The pervasive influence of social media has intensified scrutiny of their potential adverse impacts (Allcott et al., 2020).[1] In particular, political polarization is hypothesized to be exacerbated by personalized content delivery systems that reduce search costs for belief-confirming information (Tucker et al., 2018). Designed to maximize user engagement through tailored content recommendations, these feed-ranking algorithms are thought to create 'filter bubbles' that deepen political divisions (Guess et al., 2023a). The potential consequences extend beyond online echo-chambers of radicalizing content, with recent studies linking extreme 'toxic' or hateful content with increased offline violence against minorities (Bursztyn et al., 2019; Müller and Schwarz, 2021).[2]

In response to these concerns, various government bodies, including the US Senate and the Supreme Court, are actively discussing regulations targeting algorithms.[3] While these regulations are expected to yield benefits by reducing users' exposure to hateful posts, they may impose substantial costs on platforms and users. Moreover, the feedback loop between consumer preferences and algorithms may limit the effectiveness of policies aimed at minimizing socially undesirable engagement (Hosseinmardi et al., 2024). Despite the significant behavioral implications of regulating algorithms, there is little experimental evidence on the role of user preferences or the malleability of human behavior in determining engagement with extreme posts when algorithmic content curation is disabled.

To address this gap, I conducted a large-scale RCT in collaboration with a prominent vernacular social media platform (henceforth, referred to using the pseudonym 'SM') in India that has 200 million users. The experiment involved temporarily disabling personalization algorithms for over a million users. By introducing randomized content delivery in place of the platform's typical feed-ranking algorithm, I study the causal effect of algorithmic content creation on online engagement and, specifically, interactions with toxic content.

The analysis is structured along three key research questions. First, what is the effect of turning off personalization algorithms on sharing behavior with respect to toxic posts? In other words, does exposing biased users to more diverse feeds reduce the spread of harmful content? Second, what is the cost of regulating personalization algorithms for platforms? The cost to firms is measured by the reduction in time users spend on the platform, and the consequent loss in advertising revenue. Third, to what extent are socially undesirable behaviors, such as sharing posts that are hurtful towards minority groups, driven by user

[1] According to some estimates, users spend an average of 151 minutes every day on social media platforms (Statista, 2023). See also (Braghieri et al., 2022) for a range of consequences of increased social media use.

[2] Toxicity of a post is a measure of it's hatefulness, or the harm it can cause, and is defined as per Google's Perspective API. Perspective is a free API that uses machine learning to identify 'toxic' comments, where toxicity is defined as "rude, disrespectful, or unreasonable comment that is likely to make someone leave a discussion." Perspective has been adopted by organizations like The New York Times to automatically regulate and filter abusive comments, and has also been used in academic research (Jiménez Durán, 2022). See `https://perspectiveapi.com/how-it-works/` and `https://perspectiveapi.com/case-studies/`, for a comprehensive guide and applications of this API.

[3] The regulatory policies targeting algorithms that are being considered in the US are expected to impact close to 300 million social media users. 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. See also `https://shorturl.at/G6Lj4` for SCOTUS view on Texas, Florida regulation of social media moderation.

preferences?[4] Alternatively, what are the channels through which user behavior is reinforced by exposure to toxic content via recommendation algorithms?

To answer these questions, the field intervention effectively 'switched off' the personalization algorithm for close to a million treated users, and exposed them to randomly picked content instead. Treatment was randomly assigned to 0.5% of SM's 200 million users, and exposure to content in the treatment group was given purely by chance due to design of this 'random algorithm' for the treated. This enabled the identification of supply (algorithms) and demand (user preferences) factors that drive online misinformation. In particular, I study user engagement with toxic content, where post toxicity is a measure of the harm a post can cause, as per the definition of Google's Perspective API (Jiménez Durán, 2022).

The personalization algorithm for control users continued to optimize over various metrics of user satisfaction during the intervention period, in order to maximize the time users spend on the platform. SM's typical algorithm, like that of Netflix, does this by ranking the kind of content a user has engaged with in the past (Athey and Imbens, 2019). These rank-orderings place a higher 'score' on posts that the user is more likely to share, where the likelihood of sharing is estimated using machine learning models trained on a recent history of user behavior.[5] Instead, the intervention replaced these ranked lists of posts (henceforth, called content feeds) with random draws of content, for each treated user, on each day.

Social media platforms like SM often introduce some randomly drawn posts in personalized feeds to expose users to a more diverse set of content.[6] This is done to address regulator's concerns that personalized feeds tend to become echo-chambers of like-minded opinions as recommendation algorithms sort posts based on user engagement in the past (in order to maximize user satisfaction and time spent on the platform). The intervention is expected to increase the diversity of user feeds, as treated users will see fewer posts of the type they are used to consuming. To see this, note that the 'random algorithm' allocates toxic posts to treated users with uniform probabilities, while control users who engaged with toxic content in the past would have received such content with a higher probability. I analyze trade-offs between platform usage and diversification of content exposure, and show heterogeneous responses contingent on treatment intensity.

I find that the treatment reduced the total number of toxic posts viewed and shared on average, by 27% and 19%, respectively. This change in the amount of engagement with toxic content was primarily driven by users with a high proclivity towards toxic content, who were exposed to fewer toxic posts upon being treated. However, the intervention led to a 7.8% *increase* in the proportion of toxic posts shared, as the decrease in toxic shares was smaller than the decrease in total shares. That is, users' sharing behavior (with respect to toxic posts) is inelastic relative to the exposure in toxic content, which is exogenously varied in

[4]Here, sharing particularly means sharing posts off the platform, to other social media platforms like WhatsApp.

[5]This is operationalized through contextual embeddings that are obtained by factorizing engagement matrices constructed at the user-post level (Athey et al., 2021), as is delineated in greater detail in Section 3.3. The intervention was administered using algorithmically generated latent feature vectors for each user, called embedding vectors, by factorizing an engagement matrix containing all users and posts. These embedding vectors represent some abstract measure of user tastes over social media content as learned by the personalization algorithm.

[6]SM typically randomizes a small proportion of posts in a user's feed in order to maximize learning in an algorithm operating on exploration-exploitation frontier (Dimakopoulou et al., 2017).

the experiment.

Although the intervention had limited effects on users' engagement behavior with respect to particular types of content, it significantly reduced the value that consumers derive from the platform. Random post recommendation lowers exposure to anti-minority toxic content, but also lowers usage of the platform by 14.4%. Treated users viewed and shared fewer posts of any variety, and reduced the hours spent on the application by 35.2%. Therefore, regulating algorithms is expensive for platforms. On the other hand, the platform is more likely to lose users who prefer toxic content, which may result in net social gains.

These results indicate that the average users receive greater value from sharing toxic content when they encounter fewer toxic posts during the intervention. Upon being treated, they view and share fewer number of toxic posts. However, the reduction in the number of toxic posts shared is smaller (in absolute terms) than the effect on the number of toxic posts viewed as well as the total number of posts shared. This suggests that users actively seek out posts aligning with their tastes, even when such content is not readily served to them via the personalization algorithm. I also provide complementary evidence that treated users were more likely to use the 'search' feature on the platform.

I rationalize these results, and estimate the degree of malleability in user behavior, with a simple behavioral model in which **(1)** time spent on the platform is endogenous, **(2)** users' sharing behavior reflects a balance between their intrinsic preferences for sharing toxic content and the composition of their feeds, and **(3)** the status quo algorithm serves to maximize engagement. The model also enables characterization of heterogeneous responses to the intervention, due to varying treatment intensities across user types. Finally, the model provides an intuitive estimation strategy for key behavioral parameters, like the influence of content exposure on user behavior.

Using the model to structure an analysis of the data, I find an elasticity of toxic sharing with respect to toxic viewing of only 0.16. This means that user behavior is not malleable in the short run, even upon being exposed to random content for a period of one month. The implication is that users minimally update their view of the range of ideas that are acceptable in public discourse, which they learn from the automatically generated content feeds.[7] Moreover, the model generates a testable implication, that the reduction in the *number* of toxic posts shared is driven by users with a high preference for toxic content. This is in major part due to such users reducing their overall engagement with the platform. Therefore, user behavior is largely driven by pre-existing user tastes, and the intervention had a limited effect on user behavior.

These results have a number of important policy implications for regulating digital technologies to minimize harm and political polarization, especially in developing countries. I show that while the intervention significantly reduces the proportion of toxic content viewed, a large proportion of new behaviors can be predicted using baseline preferences, and not the

[7]The range of ideas that are considered acceptable in public discourse is popularly known as the Overton Window (Astor, 2019). The Overton Window essentially informs people about the social norms, about what is considered acceptable to say in public. Matthew Desmond (2023) provides an insightful definition in his book *Poverty, by America*, relayed to him by psychologist Betsy Levy Paluck, ' 'Norms license us to do things we already believe in" (Ch 8, pp 131). Here, the idea is that users learn what is okay to say in public by observing the content that is recommended to them by the algorithm. I provide survey evidence to validate this assumption in Section 5.

exposure to new and diverse information. Counterfactual simulations suggest that any intervention that reduces toxic exposure will not be as effective in changing user behavior unless users are entirely mechanical in their responses to content exposure. Model based counterfactuals also suggest that a combination of diversified and customized feeds can be used to lower the dissemination of toxic content, while minimizing the costs of losing users. This suggests a multipronged approach to making digital spaces less hostile.

This paper contributes to three strands of the literature. First, it contributes to the literature examining how new communication technologies on human behavior and welfare.[8] This includes the contributions of Nyhan et al. (2023), Guess et al. (2023a) and Guess et al. (2023b), where a team of social scientists collaborated with Meta (henceforth called, the Meta-Science studies), to experimentally study the role of their feed-ranking algorithms in driving political polarization in the run-up to the 2020 US Presidential Elections. These studies found no effect of their on-platform interventions on various measures of affective polarization, defined as per Iyengar et al. (2019).

The main findings are consistent with my results, as users who were served lower amounts of toxic content (or the users who prefer toxic content at baseline), do not seem to change their behavior in the short-run. I complement this work in the following ways. **(1)** My analysis less prone to experimenter demand and Hawthorne effects. This is because I use outcomes from the platform's administrative data, instead of exclusively employing my survey data. My work adds breadth to the existing research as I analyze preferences revealed 'in the wild' (Ang et al., 2013) through high-frequency engagement behavior of users with millions of posts on SM, in a much larger sample.[9] **(2)** I conduct my research in a large developing country with a vastly different political context than the US and EU. India, despite being the second-largest market for digital platforms, has been scarcely studied in terms of interactions with evolving technologies. Therefore, my paper adds depth to the existing literature by identifying behavioral responses among this population. **(3)** Model based counterfactuals help in evaluating effects of policies that are difficult to implement in the field.

Second, this paper relates to a rich literature studying the relationship between media bias and political polarization (DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2011; Chiang and Knight, 2011). Even though social media differs from traditional media in some key aspects (most notably, entry costs for content creators, and personalization of content), existing literature sheds light on how exposure to content (supply factors) changes

[8]The literature in economics has emphasized the effect of technology on 'offline' measures of welfare like protest participation (Enikolopov et al., 2020; Manacorda and Tesei, 2020; Enikolopov et al., 2018; Cantoni et al., 2023), voting behavior (Zhuravskaya et al., 2020; Gonzalez, 2021; Fujiwara et al., 2023), mental health outcomes (Allcott et al., 2020; Braghieri et al., 2022), and targeted attacks on vulnerable communities by non-state actors (Bursztyn et al., 2019; Müller and Schwarz, 2021). However, the behavioral responses in many studies using observational data from social media platforms may be confounded by the algorithm itself.

[9]While the offline effects on online engagement are important and have been widely studied, there is a growing recognition that time spent online itself has a profound impact on human welfare. Consequently, social scientists are increasingly interested in questions related with mobile phone usage (Björkegren, 2019), mobile money (Suri and Jack, 2016), digital addiction (Allcott et al., 2022; Aridor, 2022), online advertising (Goldfarb and Tucker, 2011; Brynjolfsson et al., 2024), biased news consumption (Levy, 2021), spread of misinformation (Acemoglu et al., 2021), radicalization (Hosseinmardi et al., 2020), and employment of surveillance technologies (Beraja et al., 2023).

important aspects of human behavior (Ferrara et al., 2012). Similarly, this literature provides a framework to understand the demand for slanted content (slanted news in particular, Gentzkow and Shapiro (2010); Martin and Yurukoglu (2017); Angelucci et al. (2024)).

This paper contributes to the media-bias literature by analyzing demand for slanted information, when supply is itself endogenously determined by a personalization algorithm. Prior work has asserted that changing social norms are key in driving behavioral responses to biased mass-media (Bursztyn et al., 2020). I build a behavioral model on a similar vein, where social media engagement is driven by utility derived from social-image concerns, while users seek to conform to the social norms practiced in one's network (Akerlof and Kranton, 2000). However, direct evidence on the mechanisms driving behavioral responses due to (social) media bias is scarce, especially on TikTok-like platforms that currently attracts a majority proportion of social media users, and the younger demographic in particular (Aridor et al., 2024). This paper attempts to bridge this gap.

Finally, I contribute to a burgeoning literature on the economics of artificial intelligence and algorithms (Acemoglu, 2021). Fairness concerns notwithstanding, there is a growing interest in the impact of black-box algorithms in a world where they have a wide variety of applications in government and industry (Rambachan et al., 2020). Examples of such human-computer interactions include online shopping for given personalized information, decision-making in the legal system, and hiring in the labor market (Goldfarb and Tucker, 2019). These interactions are especially important to study in the presence of behavioral responses that may generate substantial unanticipated consequences (Björkegren et al., 2020; Kleinberg et al., 2022; Agan et al., 2023).

I analyze the effect of algorithms (or their absence, thereof) in a context where these technologies have been introduced only recently. This is important as more and more people interact with technologies that generate customized feeds using different types of machine learning algorithms, in contexts with scant regulations. My work shows that while social media platforms aggressively use costly algorithms to retain users on the platform, users are active agents who often seek out the content that aligns with their pre-existing biases (Hosseinmardi et al., 2024).

The remaining paper is organized as follows. Section 2 provides background details of the context and the administrative, experimental, as well as survey data employed in this study. Section 3 delineates the design of the experiment, and provides descriptive statistics that motivate an interrogation of mechanisms with a structural model. Section 4 presents the main results from an empirical analysis of the experiment. Section 5 provides a simple theoretical framework that generates testable hypotheses to test in the data. This structural model also provides an estimation strategy to measure the influence of content exposure on user behavior. The resultant estimates are presented in Section 6. Finally, Section 7 concludes with a discussion of the implications of the results for technology policy.

2 Background and Data

I lay out the context of this study, and the data sources employed to understand the effects of turning off personalization algorithms on social media.

2.1 Social Media in India

According to some estimates, more than 600 million people in India use social media platforms (Statista, 2023). This makes India one of the largest markets for online platforms in the world, second only to China. Here, WhatsApp is the leading social media network, with 488 million users. 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.[10] 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.

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 latest round of the Demographic and Health Survey (NFHS-5) reports that vulnerable populations have lower access to mobile phones and internet, social media has become the most widely used platform for public discourse in India (IIPS and ICF, 2021). Moreover, these platforms have the potential to amplify marginal voices, but may also have grave consequences for minority communities (O'Byrne, 2019; Waldron, 2009). This makes the study of social media in India especially important.

Social media usage has proven to be harmful in the Indian setting, because various posts are known to have provoked instances of violence, in the form of mob lynchings, riots, and hate crimes (Banaji et al., 2019).[11] 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 (Conover et al., 2011; Barberá et al., 2015). Such echo chambers are likely to increase user exposure to more extreme and polarized view points in the digital space, possibly leading to radicalization (Gaudette et al., 2021; Huszár et al., 2022).[12].

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.[13] Consequently, there is still an urgent need for comprehensive policy interventions to tackle the challenges posed by social media algorithms,

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

[11]Details of the particular challenges faced by social media platforms in India are provided with other contextual details in Appendix A.

[12]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 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`

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

particularly in preventing the dissemination of misinformation and extremist content.

2.2 The Platform

I partner with SM, a prominent social media platform in India, to understand the effects of exposure to extreme content, due to content recommendation algorithms. I study how the nature of online interactions changes with my intervention in SM's rich online social network, comprising (close to) 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.3. Appendix A provides more details on the platform's features, and crucially, the characteristics of content that SM users typically engage with, as a result of these features.

On SM, 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 (with the English language being conspicuously absent), SM attracts a large proportion of young 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).

The control group consists of a random sample of users who were exposed to the usual ranked list of posts. Here, the ranking was determined by user behavior revealed to the algorithm in previous engagements, and I provide details of the personalization algorithm in Appendix B. SM posts, comprising image and video-based posts, are created by influencers on this platform, as most users do not create content themselves (see Figure D.3). The intervention does not affect the aggregate supply of content because less than 1% of SM's users were randomly allocated to the treatment group. Therefore, the intervention left the incentives of these star content creators unaffected.

2.3 Administrative Data

I employ user level data on individual characteristics, post level data on the content of the posts, and user-post level data on engagement with posts.

2.3.1 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.

2.3.2 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. These are elaborated upon in Section 3.

2.3.3 Post Level Data

The platform characterizes posts by broad tag genres, based on user generated hashtags.[14] 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 the tone of the posts (Handlan, 2020). 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 C, highlight the need for contextual embeddings that accurately characterize the potential harm that a post may cause. This is done next, using a combination of supervised and semi-supervised machine learning methods, developed at Google.

2.3.4 Toxicity Classification

The administrative data provides user-post level data on viewership and engagement, and exceeds 6 TB in size. 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.

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. 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."

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

Perspective provides the best known machine learning solution for toxicity detection, as it relies on transformer based/ deep learning models (most notably, BERT) and training data from millions of comments from different publishers that are annotated by ten human raters on a scale of "very toxic" to "very healthy" contributions (Fortuna et al., 2020). This mix of Supervised and Semi-Supervised Machine Learning methods makes the Perspective algorithm sensitive to context while assigning toxicity scores.

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). These models score a phrase "based on the perceived [negative] impact the text may have in a conversation."

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, because toxic posts are rare in the data. In Figure C.3, 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. I further validate this threshold in Appendix C 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 assigned by Perspective.

As has been emphasized before, the focus of this paper is on toxic posts that target India's minority Muslim population. Since Perspective is also correctly classifies homophobic and sexist content as toxic, I replicate the analysis on a subset of political and religious posts, where 93% of the toxic posts are anti-Muslim.[15] The analysis remains robust as close to 20% of the political and religious posts are toxic, and 31% of the toxic posts are political or religious in nature.

2.4 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.

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.

I also use the survey data validate assumptions I make about user's level of sophistication in interacting with new technologies, and their ability to understand the intervention. These assumptions are crucial to state the model of user behavior in Section 5.

[15]While sexist and homophobic content may also be political, they are often excluded from the Politics genre in the data. This could not be rectified because the algorithm that classifies posts into genres is proprietary.

3 Experimental Design

I collaborated with SM to design and conduct a large-scale and long-term experiment in collaboration with SM, to expose a million users to content that was randomly drawn from the corpus of 2 million posts in the Hindi Language, that is generated each day.

3.1 Sample

Out of the 200 million users on the platform, only about 1 million users were treated in the experiment. Approximately 4 million users 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 20% of the total user base, to ensure that the text analysis is accurate because Hindi is the native language of the author. Further, I include only active users in the sample, where I define active users as individuals who viewed at least 200 posts during the baseline period (15% of the experimental sample). These cuts reduce the sample size to a quarter million users, with 63041 in the treatment group, and 168773 in the control group.

Table 1 shows 70% of the users in this constructed sample of 231,814 users are male. These numbers are representative of the gender distribution of social media users in India as per NFHS-5, which reported that 41% of the women in India were excluded from the group of internet users (IIPS and ICF, 2021). Further, the average user in the sample created their account some time in 2022, seven years after the platform was launched in 2015, almost two years after TikTok was banned in India in 2020. This is also consistent with the estimated growth in internet penetration in India around 2022, as the proportion of internet users in the population increasing from 20% in 2018 to 46% in 2022 (World Bank, 2022). Figure D.1 shows that users with higher proclivity to toxic content at baseline are among the oldest users on the platform.

Table E.1 shows that an average SM user spent close to 7 hours on the platform during the baseline period of 31 days, which is lower than the estimated time spent on social media in India, 141.6 minutes each day. This means that the average user was actively consuming content from other social media platforms as well. The platform's integration with WhatsApp is a unique feature that implies that SM users are very active on WhatsApp, which is also the most popular social networking application in the country, with 488 million users out of a total of 600 million users (Statista, 2023). Therefore, the platform is representative of internet users in India who largely used SM to consume content that is suitable as WhatsApp forwards in private conversations.[16] This also has implications for the type of content that SM users consume. For example, an average user in my sample viewed 1087 posts during the baseline period, and the majority of this content belonged to the category of 'Greetings' and 'Devotion' in Figure D.1.

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

3.2 Randomization

Treatment was randomly assigned at the user level to 0.5% 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 intervention, so that their outcomes were not subject to contamination due to other AB tests/ RCTs running on the platform.

Users opt-in to be randomly assigned to the treatment for market research and AB tests, that are routinely conducted by the platform, at the time of account creation. This means that users did not know that they were part of the experiment, making the study less prone to experimenter demand and Hawthorne effects. 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 treatment assignment across the constructed sample described above. I assess balance in observable user characteristics across treatment and control. I 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, like the total number of posts viewed, or the proportion of toxic posts viewed, at baseline.

I cannot reject the hypothesis that treatment assignment was uncorrelated with user characteristics, either individually or jointly. Table 1 provides estimates for a randomly selected set of attributes, as the full set of user characteristics is too long for the page. Further, this Table shows that there is 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

The control group receives standard algorithmic recommendations generated for each user on the platform, while I intervene on the personalization algorithm generated for treated users. I briefly describe the algorithm, to enable an intuitive understanding of the intervention administered in the experiment in the next Section.

Like Netflix, user feeds on SM are usually customized according to preferences revealed via previous engagement with content using Field Aware Factorization Machines, henceforth called the FFM algorithms (Aggarwal et al., 2016).[17] The algorithm generates a vector of preference weights for each user with respect to some post attributes.[18] These vector-weights in the space of some post features are known as embeddings in the machine learning/ deep learning literature (Dell, 2024). This generates a ranking of posts for each user, and users are recommended new posts according to this order, on each day.

The personalization algorithm generates these preference weights, or embeddings vectors, in the space of some (latent) features, where the features could represent a user or post's

[17]These Deep Learning models are the most widely used algorithms in the tech industry, including social media platforms like TikTok, and the new 'For You' tab on X.

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

likeness to humorous or toxic content, for instance. I provide a general overview of how these algorithms work by employing data on recent user-post engagements, along with a simple example to fix ideas in Appendix B.

3.4 Treatment

Treated users were shown a list of posts that were *not* ranked according to user preferences, but were randomly drawn from the entire corpus of content in the user's chosen language. Effectively, the treatment randomized the probability of a post being recommended to a user.[19] Therefore, the control group views posts according to probabilities generated by the recommendation algorithm, while the treatment group views posts according to probabilities generated randomly.[20]

I demonstrate the key properties of the content distribution in the treatment group by simulating a simple recommender system that generates the probability with which a post would be assigned to a user by the personalization algorithm (see Appendix B for details). First, note that the distribution of the probability of content assignment among the treated, in Figure B.3, approximates a normal distribution centered around the average probability in the control group. This is predicted by the Central Limit Theorem (CLT), because the assignment probabilities are randomly picked for treated users on each day of the intervention period. Second, the CLT also predicts a smaller spread for the treated embeddings than the control embeddings, because the variance of treated embeddings is divided by the number of control users.

Crucially, the treatment has a greater effect on users with more extreme preferences over toxicity of the content. Figure B.5 shows that the users with preferences closer to the average user did not see large differences in assignment probabilities, and therefore the content feed, when treated. This is an important characteristic of content distribution, and is formally discussed as treatment intensity in Section 4. Contrary to expectations, the assignment of average preference weights to treated users does not bias exposure of treated users to popular content, as seen in Table B.1.

The intervention began on February 10, 2023 and continued till the end of the year (December 31, 2023). Administrative and survey data on relevant outcomes were gathered for the baseline period (December, 2022), intervention period (February to December, 2023) and post-intervention period (January-March, 2024).

[19]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 B for details.

[20]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 $2\sigma^2$. This way of administering the intervention has implications for the distribution and ranking of posts in the treatment group.

3.5 Descriptive Statistics

This paper studies the effect of 'turning-off' personalization of content feeds via recommendation algorithms, on measures of satisfaction and engagement on the platform, as well as on socially undesirable behavior. I first describe the effect of the intervention on the average user to understand the overall effect of the intervention.

The algorithm keeps users engaged online

Since the content recommendations are tailored according to user tastes, any deviation from personalized recommendation is expected to reduce user retention and engagement on the platform. Users value the recommendation algorithm because personalization algorithms decrease user's cost of searching for the content she prefers. In reducing ease of content discovery, the treatment reduces overall engagement with the platform, in terms of both the total number of posts viewed and shared, as shown in Table 2.

This shows that users gain value from the algorithm, and disengage from the platform when the algorithm is turned off. Table E.1 notes 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. An average treated user reduced the total time she spent on the platform by 2.5 hours, compared to an average control user, when the average control users spent close to 7 hours per month. Therefore, treatment reduced the time spent on the platform by 35%.

These results imply that the intervention was costly for the firm because the platform earns revenue from the time users spend on the app and the attention they pay to advertisements. In particular, Table 2 shows on average, treated users viewed 35 fewer posts when the control users viewed about 250 posts on average. Back of the envelope calculations suggest that were the intervention to be upscaled to the entire platform, the firm would have lost $45,817 in advertising revenue in the first month of the intervention.[21]

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. This is because the treatment intensity is expected to be positive for users who do not prefer toxic content, but negative for users who do.[22] 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. [23]

Figure D.2 shows that, during the first month of implementing this intervention (i.e. February 10 to March 10, 2023), the treatment group was exposed to less toxic content on

[21]The estimate was obtained using the price of INR 0.55 that SM charges advertisers per 1000 impressions, in an educational ad campaign designed by the author. The estimate was converted to USD using the exchange rate of 1 USD = 84.03 INR as on October 6, 2024.

[22]To build intuition, this is seen in the simulated recommendation algorithms for treatment and control groups in Figure B.5, where all treated users are exposed to toxic content with a uniform probability.

[23]The probability of being assigned toxic content does not necessarily equal the inverse of the number of toxic posts in the corpus. This is because the probability is determined by the cross product of user and post embeddings (detailed in Append B). As a result, the average treatment intensity does not equal zero in this experiment.

average. On average, random content delivery serves treated users with a lower number of toxic posts.

Online sharing behavior is inelastic

The average user viewed fewer toxic posts in the intervention period than in the baseline period. I expect to see a reduction in the total number of toxic posts shared because **(1)** treated users have higher search costs of seeking out toxic posts to share from, and **(2)** seeing more diverse content may change their (toxicity related) attitudes. However, the average treated user increased the fraction of toxic posts she shared for each toxic post she viewed.

Table 2 shows that while treatment reduced the proportion of toxic posts viewed by 8.7%, it increased the proportion of toxic posts shared by 7.8%. This is because even though an average SM user saw less toxic posts, and shared a smaller number of toxic posts, she shared far fewer posts of other categories. Users seem to be inelastic, and tend to seek out content that they prefer, even when the algorithm does not serve it to them readily.

Furthermore, the decrease in number of toxic shares (20%), is not as large as the decrease in number of toxic views (27%). Therefore, the elasticity of toxic sharing with respect to toxic viewing, defined as the ratio of percentage change in number of toxic posts shared to the percentage change in 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.

Treatment induces behavioral responses to seek out content

On average, the treatment effect on the ratio of toxic shares to toxic views is positive. This means that the average treated user changed her behavior in response to the intervention, undercutting the negative treatment effect on the number of toxic posts shared. To see 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.[24]

On average, I find that the exposure and disengagement effects contributed to 66% of the reduction in number of toxic posts shared. This means 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 of toxic views to toxic shares were to be less than or equal to zero, the treatment effect on the number of toxic posts shared would have been more negative.

[24]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.

An Illustration of User Behavior

While the treatment significantly decreased the number of toxic posts viewed by an average user, the reduction in the number of toxic posts shared was not as large. The descriptive evidence points to the fact that user behavior is not malleable or elastic, as sharing behavior does not change as much as the views. I illustrate the point using a simple example.

Consider a user who is served 15 posts in a day, and she shares 9 of them. If the user is served 5 toxic posts and 10 non-toxic posts, and she shares 2 toxic and 7 non-toxic posts, then the proportion of toxic posts shared is 22%.

Now, consider the treated user. The user views 2 toxic posts and 7 non-toxic posts. Suppose, she shares 1 toxic post, but shares only 3 non-toxic posts. Thus, she is disengaged from the platform and shares a total of 4 posts, instead of the 9 she would have shared, if she were not treated. Notice that she also views a smaller number of posts.

This example illustrates that even though the average user views and shares fewer toxic posts upon being treated, there is an increase in the proportion of shares that are toxic. This is because 1 out of 4 shares is toxic under treatment, meaning the proportion of shares that are toxic is 25%. On the other hand, the proportion of toxic posts shared is 22% for control users. This is true even though the user shared a lower number of toxic posts.

Treated users search more

Table E.1 shows that treated users were more likely to use the search feature on the platform. This complements the evidence on stickiness of sharing behavior, as my measure of shares includes posts accessed both through the trending feed tab, and the search tab. This finding is also consistent with the fact that treated users were more likely to view fewer posts during the intervention period, as my measure of views does not include posts accessed through the search tab.

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

4 Results: Five Facts

The intervention replaced personalization algorithms with random content delivery for 0.5% of users on a large social media platform. This helps in identifying the following five facts from the data.

Fact I: Disabling the algorithm has heterogeneous effects

The intervention assigned content according to an average user's content feed, so it reduced the amount of a particular content for users whose feed would otherwise include a lot of it (above average) and increased the amount for users whose feed would otherwise include little (below average). Therefore, treatment intensity is higher for these extreme users because their baseline exposure to toxic content was farther away from the average user's feed

To understand potential heterogeneous effects, I rank users based on their initial feed, with respect to the percentage of their feed that was toxic at baseline. This allows me to compute the effects on users who initially had low, medium, and high levels of toxic exposure. The representation of user types using baseline exposure is accurate because the personalization algorithm recommends posts according to users' past behavior.

Figure 1 shows that the treatment effect on proportion of views that are toxic is negative for users with high degree of toxic exposure at baseline (Q3 to Q5), and is positive for users with low baseline exposure to toxic content (Q1 and Q2).

Fact II: Usage declines when personalization is turned off

Users receive positive value from the personalization algorithms as they reduce the search cost of content discovery. Further, users with higher treatment intensity are expected to reduce overall platform usage more, because an average feed is farther away from their baseline feed.

Figure 2 shows that Q5 users viewed fewer posts of any type during the intervention period. Moreover, the absolute value of the effect on number of posts viewed is the larger for Q5 users than Q3 users (who have lower treatment intensity). 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.[25] However, Figure D.10 shows that there are no systematic differences in the effects on the number of times a user logged onto the platform across user types. This suggests that users largely disengaged on the intensive margin rather than on the extensive margin.

The differences in treatment intensity by user type would imply that Q1 users should also be expected to disengage with the platform. This is because users with lower proclivity to toxic content were more likely to be assigned toxic content during the intervention period, lowering the value they receive from using the platform due to this mismatch. Surprisingly, this does not seem to be the case in Figure 2. This is tackled in the discussion of the following facts.

Fact III: Elasticity of sharing toxic content is heterogeneous

The elasticity of toxic sharing is defined as the ratio of the percentage change in toxic posts shared to the percentage change in toxic posts viewed. Due to differing treatment intensities, the elasticity of toxic sharing can also be expected to be heterogeneous across user types.

Figure 3 shows that Q5 users do reduce the number of toxic posts they shared upon being exposed to fewer toxic posts. On the contrary, Q1 users, with the lowest proclivity to toxic content did not increase the number of toxic posts they shared, even though the increase in number of toxic posts they were exposed to was substantial.

[25] This may raise concerns about differential attrition in the treatment and control groups. That is, the main estimates may be biased if the type of treated users who left the platform were different from those who stayed. However, controlling for baseline engagement with toxic content and treatment status, Table E.2 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, and so, the present analysis by user types is robust to differential attrition.

Q1 users are more inelastic in sharing toxic posts with respect to toxic exposure, as the elasticity of toxic sharing is 0.08 for Q1 users, and 0.69 for Q5 users. This explains why Q1 users did not decrease the total number of posts viewed during the intervention period, as was expected for extreme users, despite high and positive treatment intensity in Figure 2. Q1 users could be seeking out non-toxic content on the platform because the total number of posts viewed (of any variety) does not decrease for these users.

Figure D.11 shows that while treated users were more likely to use the text-search feature on the platform, the treatment effect on the number of times a user searched for any content (per post viewed on the landing page) is not distinguishable across Q1 and Q5 users. While illuminating, this evidence is suggestive because of the reasons pointed out in Section 3.5.

Fact IV: Behavioral responses dampen benefits of regulations

The intervention makes it harder for treated users to discover content that would have ordinarily been recommended to them by the personalization algorithm. This may lead users to share fewer posts of the type they have higher proclivity towards, if users do not change their behavior. On the other hand, if users change their behavior, they may seek out the content they like, 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 being treated, and the treatment effect is monotonically decreasing in user type. However, relative to the decreased number of toxic posts viewed, Q5 users actually increased the proportion of toxic posts shared, as seen in Figure 4.

I claim that there is a change in behavior as users with higher proclivity to toxic content at baseline amplify their behavior by sharing a higher proportion of toxic posts they view. Figure D.7 decomposes the effect on the number of toxic posts shared, according to the empirical decomposition in Equation 1 in Section 3.5, for different user types. This shows that for users in Q5, 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.

The implication is that if this effect were 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 dampen the societal benefits of the intervention, which are given by the decrease in the total number of toxic posts shared.

Fact V: Cross-platform regulation is necessary

We have noted heterogeneity in the effects of the intervention on different types of users. However, it is unclear why users with different proclivities to toxic content at baseline respond differently to the intervention. I employ users' baseline behaviors and attributes to characterize the preferences of different user types for different types of content.

Figure D.1 shows that, irrespective of treatment status, users with the highest affinity to political and toxic content at baseline (Q5) were more likely to **(1)** have started using the platform before Q3 users, **(2)** be less active at baseline, but more active during working hours, **(3)** be male and older, **(4)** be more engaged with Romantic and Political content at baseline that Q1 users, and **(5)** be less engaged with Greetings, and Humorous content at baseline, that Q1 users.

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 more Greetings and Humorous content at baseline, which is a very particular type of content that is extremely popular on WhatsApp in India. On the contrary, Q5 users consume more 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 content users want to share on WhatsApp, in particular. 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.[26] In Figure D.14, users with the highest affinity to greetings at baseline did not disengage with the platform upon being treated. This may explain the inelasticity of Q1 users in sharing toxic content.

This is also supported by the survey evidence in Figure D.18, which shows that treated users with higher affinity to toxic content spent more time on other platforms similar to SM. On the other hand, no such trend was observed for users with lower affinity to toxic content. This links the heterogeneity in responses to the substitutability of preferred content on other platforms. In particular, users with higher proclivity to toxic content at baseline are more likely to find toxic content on other platforms, when the intervention reduced their exposure to toxic content on SM. Therefore, cross-platform regulation is required to effectively reduce the spread of toxic content.

Taking Stock

In turning off the algorithm, the intervention exposed a random set of users to an average content feed on the platform, irrespective of their baseline preferences. This led to a reduction in overall platform usage, which was costly for the platform. However, the intervention also led to a reduction in exposure to and engagement with toxic content, especially for the users with higher proclivity to toxic content at baseline. This was driven by a combination of mechanical changes in exposure to toxic content due to the intervention, endogenous responses in overall platform usage, and behavior changes.

Furthermore, there was a positive effect on the number of toxic posts viewed by non-toxic users, but there was no commensurate increase in the proportion of toxic posts shared in this category of users. These results demonstrate that user behavior with respect to toxic content is inelastic, but is even more so for users with lower proclivity to toxic content at baseline. This may be the case because Q1 users like content that is not substitutable on other platforms, while toxic users can access the content they prefer on other platforms.

Even though the intervention reduced the total number of toxic users, views and shares on the platform, the treatment led to an average increase in the proportion of toxic posts shared to toxic posts viewed. This unintended consequence is consistent with the idea that toxic users seek out the content they like, either on the platform or on other platforms.

I focus my attention on toxic users because it is not clear if the intervention resulted in a net benefit from the perspective of a benevolent social planner, who is interested in reducing

[26]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`

the spread of toxic content. First, the intervention led to a reduction in the number of toxic posts viewed and shared, but increased the proportion of toxic posts shared by toxic users, relative to the toxic posts they view. Second, the intervention displaced toxic users from the platform, but only due to substitutability of toxic content on other platforms. I quantify these trade-offs in the following section.

5 Model

Personalization algorithms have been criticized for their role in radicalizing users by creating echo chambers of like-minded individuals, who may engage with toxic content (O'neil, 2017; Zuboff, 2019). I intervened upon online social networks by replacing the personalization algorithm with a 'random algorithm.' The model organizes the empirical results to rationalize opposing effects of the intervention on the absolute and relative quantities of toxic shares. This also provides an estimation strategy to recover the behavioral parameters of users, that cannot be directly uncovered from the data.

5.1 Overview

User incentives are modeled in the spirit of Akerlof and Kranton (2000), where the user has self-image concerns. The user cares about how her action of sharing a post compares with reference to a combination of behavior among the population as well as one's own true characteristics. This provides the micro-foundations to estimate parameters of a behavioral model, where the main parameter of interest is the rate at which users update their sharing behavior when exposed to new content, that may be slanted.

The behavioral responses depend on user's exposure to content through the recommendation algorithm, which enables the user to perceive what is socially acceptable. This can aid policy instruments aimed at reducing engagement with harmful content, if diversified feeds expose users to alternative view points. For example, users who were more likely to engage with toxic posts at baseline were shown more neutral posts, and this may influence them to positively change their sharing behavior.

However, other endogenous responses, driven by fixed tastes for different types of content, are likely to reduce the time a user spend on the platform. This makes the intervention costly for the platform in terms of advertising revenues. The estimation strategy is designed to disentangle these two types of responses, and identify the influence of exposure through the algorithm.

5.2 Setup

I highlight user incentives to share toxic content, as well as the algorithm's objective to maximize user engagement.

5.2.1 Platform and Algorithm

The algorithm's objective is to optimize post assignment across various types of content to maximize user engagement with the platform. I denote the probability of the algorithm

assigning toxic and non-toxic posts as q^t and q^n, respectively. The total number of posts viewed by each user, denoted by N, is endogenously determined by the user for these given probability assignments chosen by the algorithm. In optimizing the total number of posts viewed or the total time spent on the platform, the algorithm is maximizing the attention paid to various advertisements hosted on the platform.[27]

5.2.2 User

Consider a social media user who chooses S^r posts of type-r to share out of the total N^r posts of type-r posts viewed, where $r \in \{t \text{ (toxic)}, n \text{ (non-toxic)}\}$. Sharing is assumed to be costly for users, limiting the total number of posts she can share to $S = S^t + S^n$. Effectively, the user chooses the proportion of posts shared that are toxic as, $s^t = S^t/S$. She also picks the utility maximizing number of posts to view, $N = N^t + N^n$, for given assignment probabilities q^t and q^n determined by the algorithm. The premise is that choosing s^r generates some consumption utility. The user obtains this consumption utility from both viewing and sharing posts.

However, users in this model also have a taste for conformity, and derive public recognition utility from sharing posts that are closer to the average user's action (Butera et al., 2022). I assume that users' learn about society's tastes for toxic content from the content feeds that the algorithm curates for them (Salganik et al., 2006). That is, users are assumed to be sophisticated, so that they infer their social groups tastes from the content they view on the platform. I provide survey evidence to support this assumption.[28]

The user's objective is to maximize utility she derives from sharing different types of posts.

$$\begin{aligned}\max_{s^t,S,N} u(s^t, N; q^t, \beta, \alpha) = & \underbrace{\beta N - \alpha(N-S)^2 - \eta S^2}_{\text{consumption utility}} \\ & - \delta S \quad \underbrace{(1-\theta)\left(\log\left(\frac{s^t}{p^t}\right)\right)^2}_{\text{disutility of deviating from own type}} \\ & - \delta S \quad \underbrace{\theta\left(\log\left(\frac{s^t}{q^t}\right)\right)^2}_{\text{disutility of deviating from societal tastes}}\end{aligned}$$

where, β is users' intrinsic motivation, as in Bénabou and Tirole (2006), to view content or spend time on social media. α parameterizes users' intrinsic motivation to share posts of

[27]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. This greatly simplifies the analysis, because 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 to analyze the algorithm itself.

[28]Figure D.16 shows that most users in the experimental sample believed that their engagement activity affects the feeds of other similar users on the platform. Further, these responses were not primed by the treatment itself because treated users were not more likely to notice network effects.

type r, and p^t is the user's preference parameter with respect to toxic content. The utility function implies that even when users receive positive utility from viewing another post (β), she incurs some disutility if the additional posts she views is not shareable, according to her tastes (α). $\eta > 0$ results in convex cost of increasing total number of posts shared.[29]

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 others do (q^t) and their own tastes (p^t) (Mullainathan and Shleifer, 2005). In this action-signalling model, the user wants to conform with this average behavior, by choosing s^t that implies individual tastes close to society's preferences (Becker, 1991; DellaVigna et al., 2012; Dupas et al., 2024).

Therefore, $\theta \in [0, 1]$ is the weight users put on their perception of society's tastes for type-t content. Users are assumed to update their behavior in line with their perception of norms, at some rate θ. Then, θ measures 'influence' on account of exposure to the algorithmically generated content feed. This is the main behavioral parameter of interest, that is estimated in the model.

5.3 Model Predictions

The strategic interaction between the users and the algorithm unfolds according to the timing described in Appendix G, where I also solve for the subgame perfect equilibrium. Consider two time periods in this model, so that $\tau = 0$ represents the baseline and $\tau = 1$ denotes the intervention period. The algorithm assigns toxic posts to each user i at time τ with probability $q^t_{i,\tau}$, with $q^n_{i,\tau} = 1 - q^t_{i,\tau}$.

The model allows a characterization of user types in terms of baseline exposure to toxic content. This is because Lemma G.5 shows that the equilibrium assignment probabilities exactly equal respective users' tastes for such content, i.e. $q^t_{i,0} = p^t_i$ for all i at baseline $\tau = 0$. Therefore, I analyze the main outcomes separately for users in different quantiles of baseline exposure to toxic content, or the distribution of proclivities to toxic content.

The intervention is symbolically represented as the average of assignment probabilities in the control group ($\bar{q_1}^t = \mathrm{E}[q^t_{i,1}|D_i = 0]$).[30] Therefore, I represent the treatment effect as the change in an outcome, with respect to changes in the exogenous probabilities assigned under treatment, all else equal. That is, the treatment effect is the partial derivative of an outcome with respect to $\bar{q}^t$.

The comparative statics deviate from the experiment by only treating users with higher proclivity to toxic content, that is users with $p^t_i > \bar{q}^t$.[31] This is because the previous section demonstrated that the treatment increased the number of toxic posts viewed by treated users with the lowest proclivity to toxic content. Such an effect is not desirable, either from the platform's (Q1 users are not seeing what they like), or the planner's perspective (Q1

[29] This is the cost of sharing too many posts may reduce the attention paid to a user's shared posts.

[30] This is because these probabilities are picked uniformly at random, each day during the intervention period, from the set of all possible assignment probabilities in the control group. Then, the Law of Large Numbers ensures probability convergence to the average assignment probability in the control group.

[31] Such a targeted policy could not be implemented in the field experiment. This is because the platform does not want to target users based on their proclivity to toxic content, and does not want to take a position on what content is toxic, in order to remain politically neutral. The implementation of the experiment was constrained by these considerations.

users are seeing more toxic content). The targeted policy allows a more direct approach to reducing toxic exposure for the most toxic users, and is still empirically validated by focusing on the most toxic users (Q3 to Q5).

Prediction 1: Treatment intensity

Users in Q5 have a higher taste for toxic content, and are expected to be most affected by the treatment. This is due to the largest reduction in exposure to toxic content for treated users in this group. Then, the model prediction on treatment intensity, or the proportion of posts viewed that are toxic, under the alternative intervention, is stated in the following proposition.

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

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

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

Proof. $\bar{q}^t - q_{i,\tau}^t < 0$, for more toxic users 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 completed the proof. □

Figure 1 shows that the treatment effect on the number of toxic posts viewed is negative for toxic users. Since the design of the counterfactual policy is different from the intervention that was actually implemented, I only analyze outcomes for users in Q3 to Q5 to validate the model empirically.

Prediction 2: Number of posts viewed

The model predicts that, toxic users (or users with higher baseline exposure to toxic content) view fewer posts of any variety upon being treated.

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

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

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.

Proof. In Appendix H. □

Intuitively, this is because the treatment exogenously lowers the probability of being assigned toxic content to the control mean, so that small increases from $\bar{q}^t$ bring this probability closer to the user's true taste for toxic content, p^t.

In Figure 5, I simulate the model's predictions with respect to total posts viewed, under two regimes: treatment and control. The control users continue viewing the optimal number of posts in equilibrium, but treated toxic users are shown to view fewer toxic posts, and therefore choose to view a lower number of posts in total. Figure 2 verifies that the number of posts viewed is monotonically decreasing in user tastes for toxic content.

Prediction 3: Proportion of posts shared that are toxic

The model makes pertinent predictions on the proportion of toxic posts shared, $S^t_{i,\tau}/S_{i.\tau}$. This is informative about benefits of the intervention, in terms of discouraging toxic behavior (total number of toxic posts shared, $S^t_{i,\tau}$), relative to the costs borne by the platform, in terms of losing engagement (total number of posts viewed, $N_{i,\tau}$, and shared $S_{i,\tau}$).

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

$$\frac{\partial^2 s^t_{i,\tau}}{\partial p^t_i \partial \bar{q}^t} \geq 0 \tag{3}$$

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 with higher proclivity to toxic content.

Proof. In Appendix H. □

Figure 5 shows that the treatment effect on the proportion of toxic posts shared is negative for users with higher proclivity to toxic content. Intuitively, this is because marginal changes in the proportion of toxic posts viewed due to the treatment, only changes sharing behavior by an order of θ, through the channel of influence from exposure.

This implication is tested in the data with Figure D.12. Here, the treatment effect on proportion of shares that are toxic is negative for users in Q5.

Prediction 4: Mechanical Effects

These results demonstrate how different type of users respond differently to the treatment. If users were mechanical, they would all have constant effects 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, because they seem to be larger for more toxic users. This means that users put a positive weight on the new information they receive, when making sharing decisions. The users are considered to be 'behavioral' in this sense.

The model predicts mechanical behavior if and only if the influence parameter, θ equals 0 for mechanical users. Before estimating θ in the behavioral model, I show that this parameter is, in fact, non-trivial.

Proposition 4. *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.

Proof. In Appendix H. □

In other words, 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. I reject 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 will show in estimating this model.

5.4 Estimation

The model provides an estimation strategy for the rate at which users update their sharing behavior, θ, in line with what they think is socially acceptable.

5.4.1 Measurement

The data consists of measurements on number of toxic and non-toxic posts viewed and shared by each user, during the intervention period ($\tau = 1$) and the baseline period ($\tau = 0$). I measure sharing probabilities during the intervention period as, $s^r_{i,1} = \frac{S^r_{i,1}}{S_{i,1}}$, for $r \in \{t, n\}$.

I imperfectly measure users' innate preferences with their sharing behavior at baseline ($\tau = 0$), i.e. $p^t_i \equiv s^t_{i,0}$ and $p^n_i \equiv s^n_{i,0}$. This is because user behavior at baseline is said to be in equilibrium, with $q^t_{i,0} = s^t_{i,0} = p^t_i$. Therefore, Lemma G.1 implies that $s^t_{i,0} = p^t_i$ in equilibrium.

Next, I proxy user's probability of being assigned toxic content with the proportion of toxic posts viewed, i.e. $q^t_{i,\tau} \equiv v^t_{i,\tau}$. This is important because the probability of being assigned toxic content is not observed during the first month of the intervention, and therefore, I assume that a user views all the posts she is assigned by the algorithm. I later relax this assumption, and employ a measurement error correction to account for the fact that users only view a proportion of the posts they are assigned. Then, the equilibrium sharing function in Lemma G.1 is represented as functions of sharing behavior at baseline, and views during the intervention period.

$$\log s^t_{i,1} = (1 - \theta) \log s^t_{i,0} + \theta \log v^t_{i,1} + \mu \log w^t_i \tag{4}$$

where, w^t_i is a taste-based, and time invarying preference shock for sharing a post. The interpretation of the main parameter of interest, θ, as the influence of exposure, 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 equations (4) due to certain features of the experiment's design, which are elaborated upon in Appendix I.1. A steady state condition is used to identify θ.

5.4.2 Steady State

This system is said to be in steady state when the probabilities of assigning toxic and non-toxic posts ($[q^t_{i,\tau}, q^n_{i,\tau}]$), as well as the probabilities of sharing toxic and non-toxic posts ($[s^t_{i,\tau}, s^n_{i,\tau}]$) are stable over time. The steady state condition is also the identifying condition as it states that 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. This allows for the identification of the updating parameter θ.

That is, in the sample of treated users, θ is identified when the following assumption is satisfied,

$$s^t_{i,0} = s^t_{i,1} = \left(v^t_{i,1}(\bar{q^t})\right)^{\theta}\left(w^t_i\right)^{\mu}\left(s^t_{i,0}\right)^{1-\theta} \tag{A2}$$

I test the validity of this assumption using the sample of control users in Appendix I.6. I consider log-likelihoods due to skewness in the distribution of shares, to arrive at the following Proposition.

Proposition 5. *For some updating parameter θ, and treated user i, the change in ratio of toxic-shares to non-toxic shares from the baseline is a function of the log-odds ratio of the proportion toxic posts viewed at baseline. That is,*

$$\log\left(\frac{s^t_{i,1}}{s^n_{i,1}}\right) - \log\left(\frac{s^t_{i,0}}{s^n_{i,0}}\right) = (1+\theta)\log\left(\frac{\bar{q^t}}{\bar{q^n}}\right) - \theta\log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right) \tag{5}$$

where, $\bar{q^t}$ and $\bar{q^n}$ are constant.

Proof. In Appendix H. □

Intuitively, the constructed outcome in Proposition 5 accounts for unobserved differences in preference for sharing any type of content. As a result, θ can be identified using the relationship between *difference in shares* (from $\tau = 0$ to $\tau = 1$) and *levels of views at baseline* ($\tau = 0$) in the sample of treated users.

6 Estimates

I flexibly fit the probability of sharing toxic content with respect to the difference in the proportion of toxic posts and non-toxic posts viewed at baseline (as an approximation to the relevant log-ratios, henceforth referred as simply the difference). The function in Figure D.21a approximates a linear relationship.

6.1 OLS Estimates

Table E.6 shows that the estimated effect of a 1% decrease in the proportion of toxic posts viewed during the intervention period decreases the proportion of toxic posts shared by $\hat{\theta}\% = 0.08\%$. This also demonstrates stickiness in user behavior, as the elasticity in sharing behavior with respect to baseline sharing behavior, $(1 - \hat{\theta})$ is close to 0.92.

These estimates support the claim that user behavior is not malleable, and is largely determined by user preferences at baseline. However, these OLS estimates of θ are likely to suffer from attenuation bias. This is because the proportion of toxic or non-toxic posts viewed are measured with error, due to the fact that treated users sample a fraction of posts to view, after they are assigned content randomly. I use an IV strategy, outlined in Appendix I.3, to correct for this possibility.

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. Then, the IV estimates, in Column (2) of Table 3, indicate that the measurement error indeed attenuated the main OLS estimates. The first column in Table 3 shows the strength of the first stage in the IV specification. The corrected estimate shows that a 1% reduction in exposure to toxic content reduces engagement with toxic content by 0.16%.

The IV strategy provides the preferred estimate of the elasticity of sharing toxic content with respect to viewing toxic content. More standard approaches for correcting classical linear measurement error provide larger estimates of θ. This first stage also serves as a measure of reliability for the main outcome variable, in STATA's in-built measurement error correction program, whose result is shown in column (3).

The lack of flexibility in user responses has serious implications for designing regulatory policies aiming to reduce engagement with toxic content in digital spaces. My results indicate that the elasticity of the odds of sharing toxic content, with respect to exposure at baseline is more than 0.84. This leaves little room for policy instruments to alter sharing behavior through reduced exposure to toxic content.

Table 3 shows that user behavior significantly depends on pre-existing behaviors or preferences. This means that while users update their behavior in line with new information that they are exposed to, a significant part of their behavioral responses are sticky as they depend on user behavior at baseline. I perform a series of checks to validate these structural estimates in Appendix I.4.

6.3 Model Based Counterfactuals

I estimate the model parameters to predict the effect of reducing exposure to toxic content by randomizing the feed of a subset of users. This is done 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. The simulated moments are generated using the model with θ set at 0.16. The calibration exercise for various model parameters detailed in Appendix I.5.

6.3.1 Alternative Behavioral Assumptions

I simulate the effects of the counterfactual intervention, which targets more toxic users to randomize their feed, under different assumptions on user behavior. I use the calibrated model parameters to construct counterfactual distributions of 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.22 shows that for malleable users with $\theta = 1$, the percentage change in number of toxic posts shared is decreasing user in type. This is because more toxic users viewed fewer toxic posts, and are more likely to be influenced when $\theta = 1$. The decrease in number of toxic posts shared is larger in absolute terms when users are fully malleable ($\theta = 1$), than when the users behave according to the observed degree of malleability, $\theta = 0.16$. The case of mechanical users ($\theta = 0$) confirms that treatment leads to no change in the number of toxic posts shared, or in its constituent parts.

6.3.2 Treatment Effect Decomposition

Next, for the calibrated model parameters, including $\theta = 0.16$, I decompose the treatment effect on the total number of toxic posts shared into two channels: **(1) Engagement:** the change in engagement with the platform, or the number of posts of any kind, that were viewed and shared by treated users, and **(2) Influence:** the change in the probability of sharing toxic content, given the number of posts shared. Simply put,

$$S^t_{i,1} = N_{i,1} \cdot \frac{S_{i,1}}{N_{i,1}} \cdot s^t_{i,1}$$

$$\implies \%\text{ change in } S^t_{i,1} = \underbrace{\%\text{ change in } N^t_{i,1} + \%\text{ change in } \frac{S_{i,1}}{N_{i,1}}}_{\text{change in engagement}} + \underbrace{\%\text{ change in } s^t_{i,1}}_{\text{change in influence}}$$

For θ estimated at 0.16, I simulate the effect of exogenously changing $\bar{q}^t$ for different types of users, and analyze how these two components of the treatment effect change the number of toxic posts shared. Figure D.23a shows that the treatment effect on total number of posts viewed and shared reduces the treatment effect on the number of toxic posts shared. For this low value of the influence parameter, the decreasing effect on the probability of sharing toxic content seems to be largely driven by the change in engagement, especially for more toxic users.

In this counterfactual decomposition, the behavioral response, represented by s^t, is surely decreasing in user toxicity. However, the decrease in N is even larger, with increasing toxicity in user-type. Note that, when users are fully malleable with $\theta = 1$, as in Figure D.23c, the treatment effect in number of toxic posts shared in entirely driven by the change in proportion shares that are toxic, s^t. 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.

6.3.3 Counterfactual Policies

Social media platforms frequently try to diversify user feeds, by randomizing a portion of the posts that users see. Randomizing part of the feed makes for good policy because the platform algorithm can learn about changing, and sometimes inconsistent, user preferences, from their engagement with randomly served content (Kleinberg et al., 2022). 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, and continuously learning about their preferences (Zhan et al., 2021). 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 outcome, number of toxic posts shared, and its component parts, under different mixes of algorithmic and random feeds in Figure 6. This shows that even when 60% of the feed is randomized, the effect on toxic sharing for toxic users is driven by the influence effect, or the change in the probability of sharing toxic content, given the number of posts shared (s^t). However, the social gains in terms of reduction in toxic sharing are fairly limited. On the other hand, if at least 80% of the feed is randomized, the effect on toxic sharing for toxic users is driven by the engagement effect. This counterfactual exercise shows that a planner can optimally choose a degree of randomization, to balance the trade-off between user engagement with social media platforms, and the dissemination of toxic content.

7 Conclusion

Content recommendation algorithms are often accused of boosting engagement with misinformation, like hate speech, on social media platforms, all around the world (Pariser, 2011). Personalization algorithms have also been linked to hate-crimes and politically motivated violence. It is, however, unclear if these harms are propagated solely due to algorithmic exposure to toxic content, or by users' pre-existing preferences over such content.

This paper studies the role of user preferences and personalization algorithms in driving engagement with extreme content. I conducted a large-scale randomized evaluation of such algorithms by effectively turning it off for one million treated users, on a popular TikTok-like platform in India. This ensures that the content users are exposed to during the intervention is uncorrelated with their preferences and therefore, distinguishes the influence of technology. I examine whether the content presented by algorithms substantially impacts user choices, or if, conversely, users would 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 proportion of toxic posts shared with respect to toxic posts viewed. I developed a behavioral model to rationalize these results, and estimated it to find that the algorithm's influence on user behavior is relatively limited. While 55-60% of the behavioral response is due to changes in engagement, the remaining 40% is attributed to the influence of the content viewed.

This paper analyzed 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. My results have important implications for regulations in a context where most users are on WhatsApp, and the spread of misinformation cannot be checked by governments, due to end-to-end encryption. Further, the current 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 8,000 users who were part of the experiment. The broader implications of this intervention, on mental health outcomes 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.

My findings have important implications for policymakers looking to regulate platforms that employ new, and seemingly opaque technologies. This is because the total effect of an intervention that decreases user exposure to toxic content is, in part, determined by behavioral responses of users. Therefore, any technological regulations that stipulate that a problematic piece of content be removed from social media must necessarily take into account these behavioral effects. This would help in estimating the difference between intended and actual effects of the policy, before it is implemented.

While existing policy frameworks are inadequate to address the challenges highlighted in this paper, achieving consensus on new regulations is difficult due to concerns about free speech and privacy. This issue has received significant amount of media attention as the failure to prevent incitement to violence has brought social media platforms, and their algorithms, under scrutiny. This was especially true in the case of the Capitol Hill riots on January 6, 2021 in Washington DC,[32] as well as the recent indiscriminate killings in Myanmar and Ethiopia.[33] In a congressional hearing on social media's role in extremism in March 2021, Rep. Mike Doyle (Pa.) addressed chief executives of Google, Facebook, and Twitter, "the power of this technology is both awesome and terrifying, and each one of you has failed to protect your users and the world from the worst consequences of your creations."[34] This paper contributes to the ongoing debate on the role of technology in shaping user behavior, and argues for the need for a multipronged approach to regulate digital commons.

References

S. Abbott et al. *Understanding analysis*, volume 2. Springer, 2001.

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

D. Acemoglu, A. Ozdaglar, and J. Siderius. A model of online misinformation. Technical report, National Bureau of Economic Research, 2021.

A. Y. Agan, D. Davenport, J. Ludwig, and S. Mullainathan. Automating automaticity: How the context of human choice affects the extent of algorithmic bias. Technical report, National Bureau of Economic Research, 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.

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

[32] See `https://www.nytimes.com/2021/03/25/business/jack-dorsey-twitter-capitol-riot.html`

[33] See `https://bit.ly/3smUoUE`

[34] See `https://www.washingtonpost.com/technology/2021/03/25/facebook-google-twitter-hearing/`

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

C. S. Ang, A. Bobrowicz, D. J. Schiano, and B. Nardi. Data in the wild: Some reflections. *Interactions*, 20(2):39–43, 2013.

C. Angelucci, M. Gutmann, and A. Prat. Beliefs about political news in the run-up to an election. Technical report, National Bureau of Economic Research, 2024.

G. Aridor. Drivers of digital attention: Evidence from a social media experiment. 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.

M. Astor. "how the politically unthinkable can become mainstream". *New York Times*, 2019. URL `https://www.nytimes.com/2019/02/26/us/politics/overton-window-democrats.html`.

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.

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.

R. Bénabou and J. Tirole. Incentives and prosocial behavior. *American economic review*, 96 (5):1652–1678, 2006.

M. Beraja, A. Kao, D. Y. Yang, and N. Yuchtman. Ai-tocracy. *The Quarterly Journal of Economics*, 138(3):1349–1402, 2023.

D. Björkegren. The adoption of network goods: Evidence from the spread of mobile phones in rwanda. *The Review of Economic Studies*, 86(3):1033–1060, 2019.

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

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, G. Egorov, R. Enikolopov, and M. Petrova. Social media and xenophobia: evidence from russia. Technical report, National Bureau of Economic Research, 2019.

L. Bursztyn, G. Egorov, and S. Fiorin. From extreme to mainstream: The erosion of social norms. *American economic review*, 110(11):3522–3548, 2020.

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.

D. Cantoni, A. Kao, D. Y. Yang, and N. Yuchtman. Protests. Technical report, National Bureau of Economic Research, 2023.

J. Chen and J. Roth. Logs with zeros? some problems and solutions. *The Quarterly Journal of Economics*, page qjad054, 2023.

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

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

M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini. Political polarization on twitter. In *Proceedings of the international aaai conference on web and social media*, volume 5, pages 89–96, 2011.

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.

M. Dimakopoulou, Z. Zhou, S. Athey, and G. Imbens. Estimation considerations in contextual bandits. *arXiv preprint arXiv:1711.07077*, 2017.

P. Dupas, M. Fafchamps, and L. Hernandez-Nunez. Keeping up appearances: An experimental investigation of relative rank signaling. Technical report, National Bureau of Economic Research, 2024.

R. Enikolopov, M. Petrova, and K. Sonin. Social media and corruption. *American Economic Journal: Applied Economics*, 10(1):150–174, 2018.

R. Enikolopov, A. Makarin, and M. Petrova. Social media and protest participation: Evidence from russia. *Econometrica*, 88(4):1479–1514, 2020.

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

E. L. Ferrara, A. Chong, and S. Duryea. Soap operas and fertility: Evidence from brazil. *American Economic Journal: Applied Economics*, 4(4):1–31, 2012.

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.

T. Fujiwara, K. Müller, and C. Schwarz. The effect of social media on elections: Evidence from the united states. *Journal of the European Economic Association*, page jvad058, 2023.

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

T. Gaudette, R. Scrivens, G. Davies, and R. Frank. Upvoting extremism: Collective identity formation and the extreme right on reddit. *New Media & Society*, 23(12):3491–3508, 2021.

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.

M. Gentzkow, B. Kelly, and M. Taddy. Text as data. *Journal of Economic Literature*, 57 (3):535–74, 2019.

A. Goldfarb and C. Tucker. Online display advertising: Targeting and obtrusiveness. *Marketing Science*, 30(3):389–404, 2011.

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

R. M. Gonzalez. Cell phone access and election fraud: evidence from a spatial regression discontinuity design in afghanistan. *American Economic Journal: Applied Economics*, 13 (2):1–51, 2021.

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.

A. Handlan. Text shocks and monetary surprises: Text analysis of fomc statements with machine learning. *Published Manuscript*, 2020.

T. Hastie, R. Tibshirani, and M. Wainwright. Statistical learning with sparsity. *Monographs on statistics and applied probability*, 143(143):8, 2015.

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

F. Hayashi. *Econometrics*. Princeton University Press, 2011.

H. Hosseinmardi, A. Ghasemian, A. Clauset, D. M. Rothschild, M. Mobius, and D. J. Watts. Evaluating the scale, growth, and origins of right-wing echo chambers on youtube. *arXiv preprint arXiv:2011.12843*, 2020.

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.

S. Iyengar, Y. Lelkes, M. Levendusky, N. Malhotra, and S. J. Westwood. The origins and consequences of affective polarization in the united states. *Annual review of political science*, 22(1):129–146, 2019.

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.

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.

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.

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

J. Ludwig and S. Mullainathan. Machine learning as a tool for hypothesis generation. *The Quarterly Journal of Economics*, 139(2):751–827, 2024.

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.

R. Mukherjee. Mobile witnessing on whatsapp: Vigilante virality and the anatomy of mob lynching. *South Asian popular culture*, 18(1):79–101, 2020.

S. Mullainathan and A. Shleifer. The market for news. *American economic review*, 95(4): 1031–1053, 2005.

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.

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.

W. I. O'Byrne. Educate, empower, advocate: Amplifying marginalized voices in a digital society. *Contemporary Issues in Technology and Teacher Education*, 19(4):640–669, 2019.

C. O'neil. *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown, 2017.

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

A. Rambachan, J. Kleinberg, S. Mullainathan, and J. Ludwig. An economic approach to regulating algorithms. Technical report, National Bureau of Economic Research, 2020.

M. J. Salganik, P. S. Dodds, and D. J. Watts. Experimental study of inequality and unpredictability in an artificial cultural market. *science*, 311(5762):854–856, 2006.

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

Statista. Daily time spent on social networking by internet users worldwide from 2012 to 2023 (in minutes) statista. Technical report, 2023. URL `https://www.statista.com/statistics/433871/daily-social-media-usage-worldwide/`.

T. Suri and W. Jack. The long-run poverty and gender impacts of mobile money. *Science*, 354(6317):1288–1292, 2016.

N. Thakral and L. T. Tô. *When Are Estimates Independent of Measurement Units?* Boston University-Department of Economics, 2023.

K. E. Train. *Discrete choice methods with simulation*. Cambridge university press, 2009.

J. A. Tucker, A. Guess, P. Barberá, C. Vaccari, A. Siegel, S. Sanovich, D. Stukal, and B. Nyhan. Social media, political polarization, and political disinformation: A review of the scientific literature. *Political polarization, and political disinformation: a review of the scientific literature (March 19, 2018)*, 2018.

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.

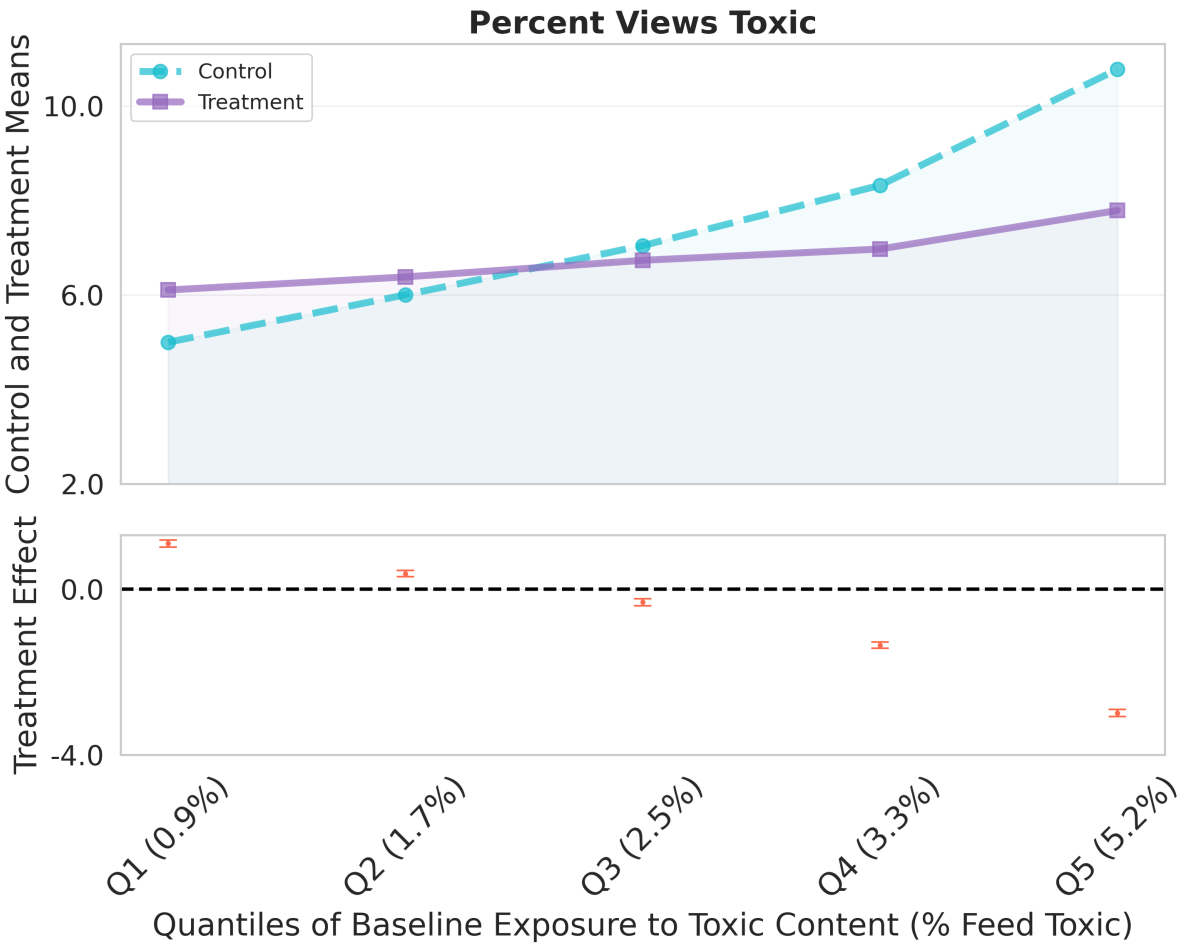
R. Zhan, V. Hadad, D. A. Hirshberg, and S. Athey. Off-policy evaluation via adaptive weighting with data from contextual bandits. In *Proceedings of the 27th ACM SIGKDD Conference on Knowledge Discovery & Data Mining*, pages 2125–2135, 2021.

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

S. Zuboff. *The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power*. PublicAffairs, New York, 2019.

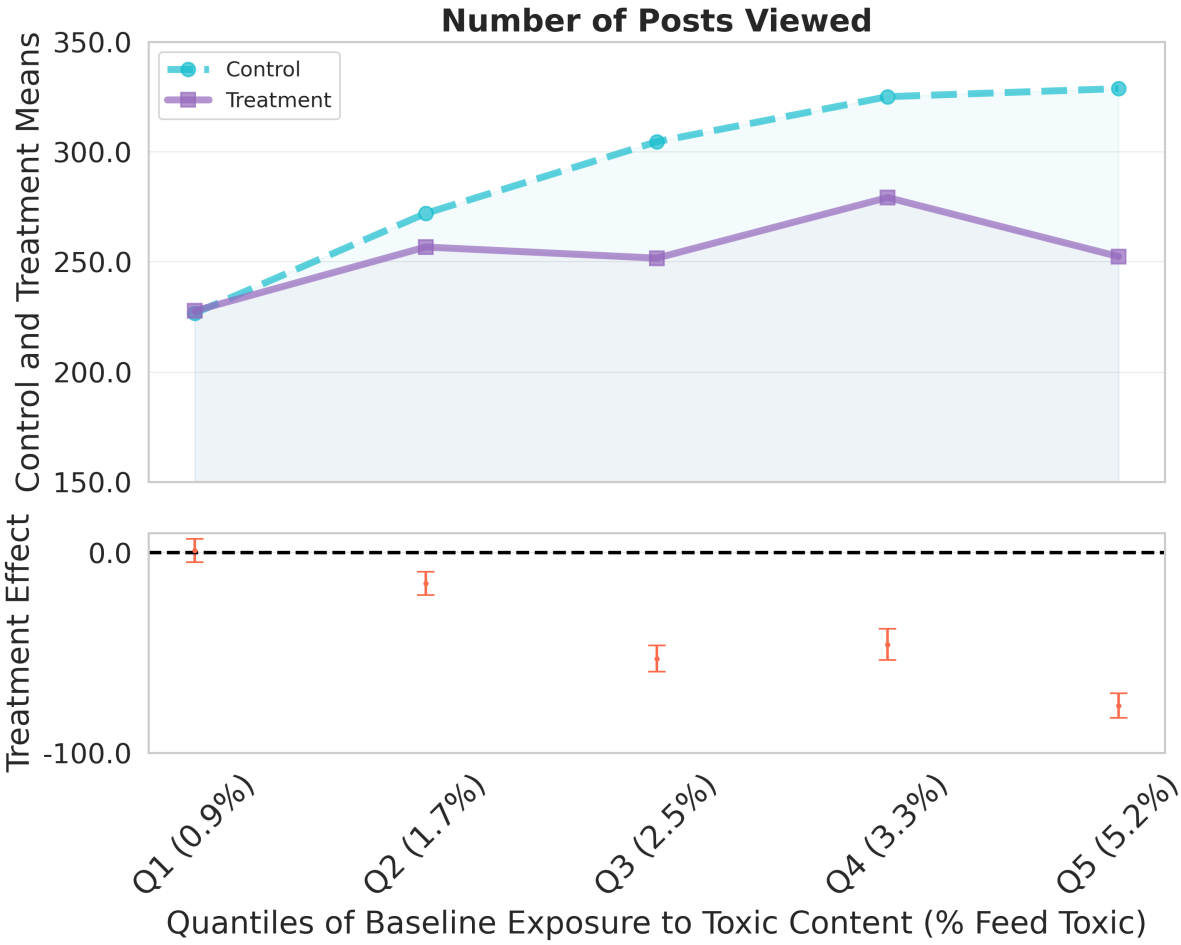
Tables and Figures

Figure 1: Treatment intensity by user type



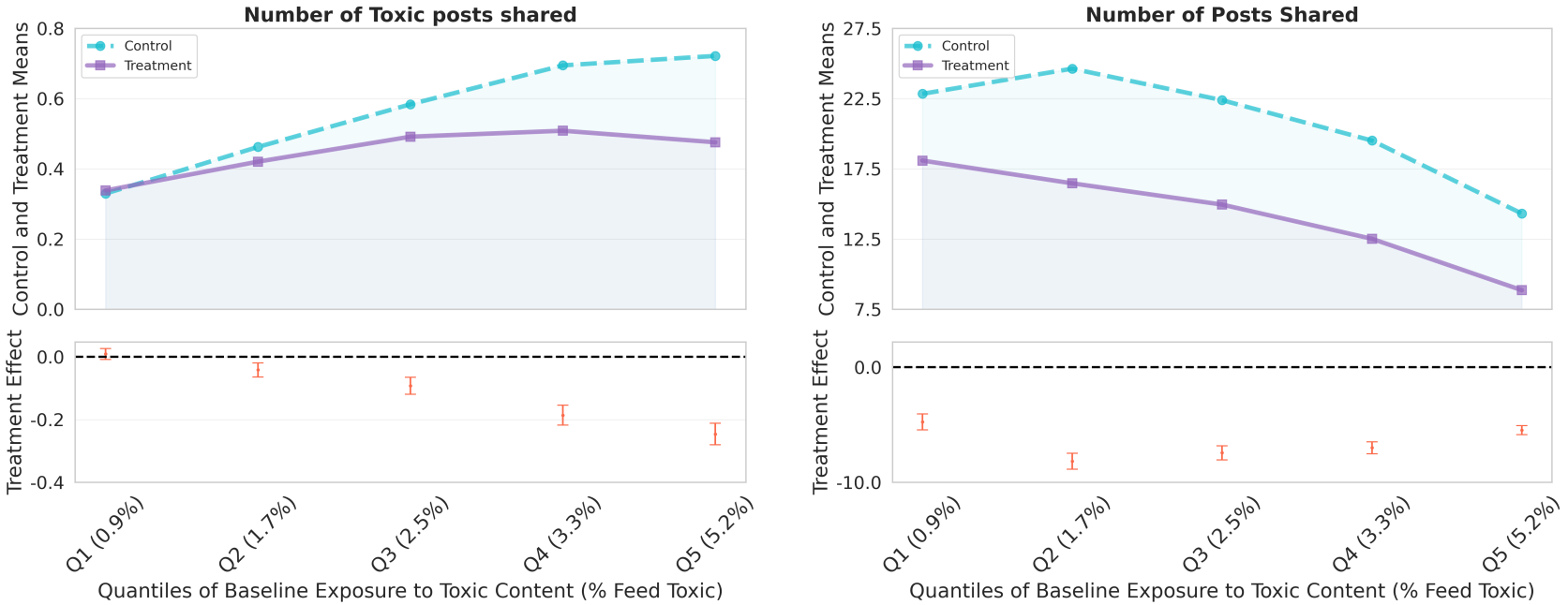
Notes: This figure shows that the treatment effect on the proportion of posts viewed that are toxic is monotonically decreasing in the exposure to toxic posts at baseline. This effect is positive for users in Q1, and negative for users in Q5, and is formalized as the treatment intensity in this experiment. The model formally characterizes user types by the proportion of toxic posts viewed at baseline, and predicted that the treatment effect on the number of toxic posts viewed is negative and larger (in absolute terms) for 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 with robust standard errors.

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



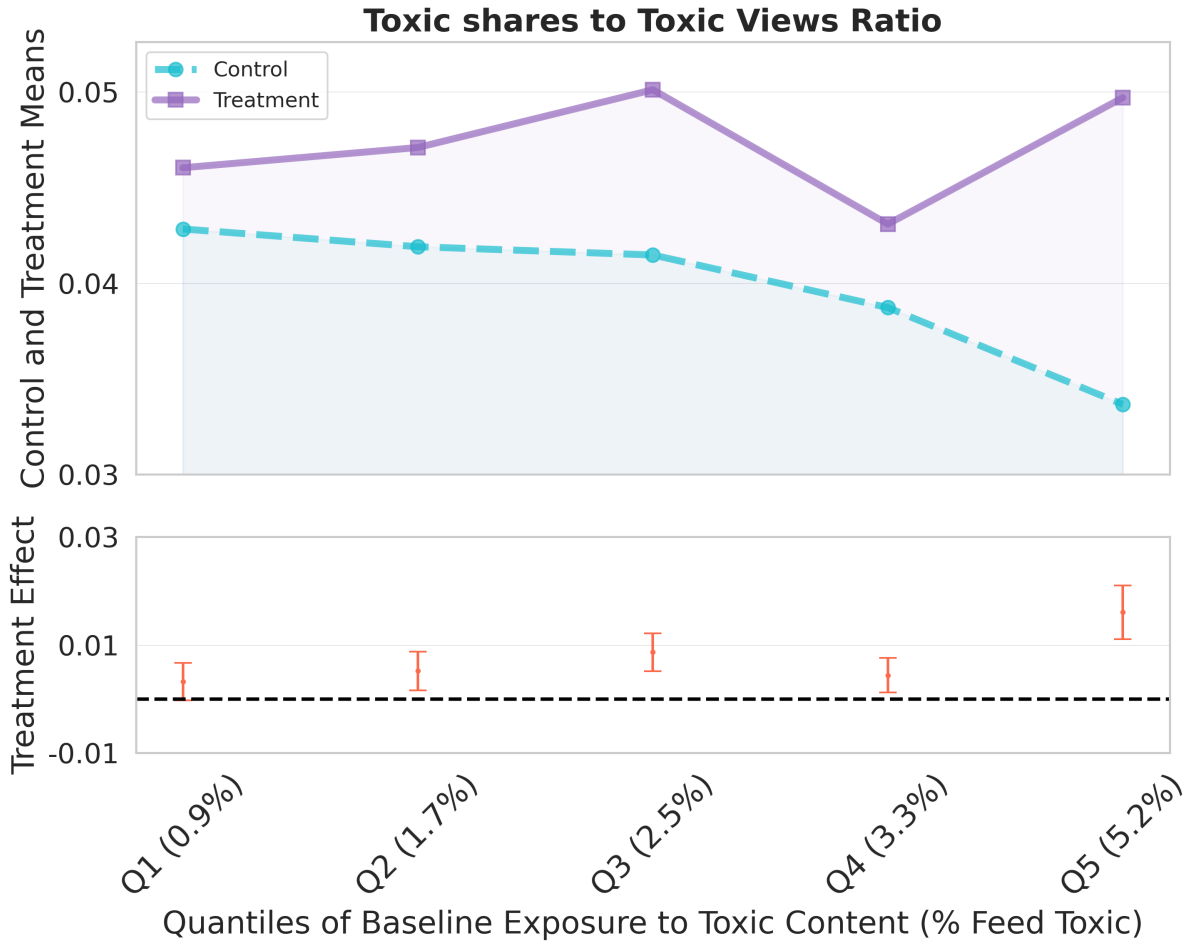
Notes: This Figure shows that he total number of posts viewed also changes by treatment status and user type. In fact, the treatment effect on the total number of posts viewed is larger (in absolute terms) for more toxic users. This is because of **(1)** lower exposure to toxic content, and **(2)** the disengagement effect. 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 with robust standard errors.

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



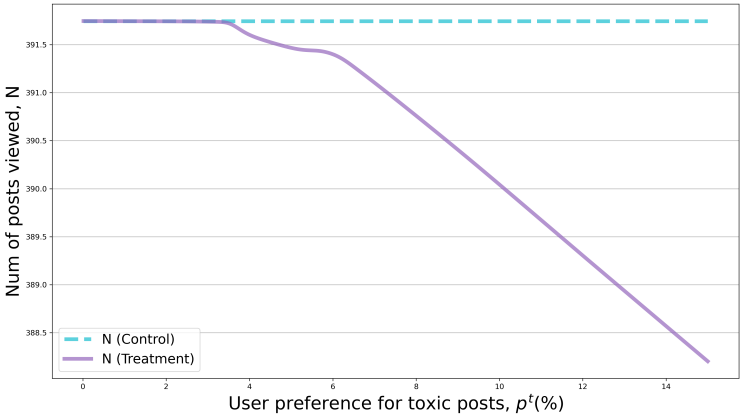
Notes: This Figure shows that the treatment effect on the number of toxic posts shared is negative for toxic users (Q3 to Q5), but is statistically insignificant (yet, positive) for users in Q1. This shows that even as there is a decrease in the number of toxic posts shared by toxic users, these users disengage with the platform by sharing fewer posts overall. From this Figure, it is therefore unclear if toxic users share fewer toxic posts because they are exposed to less toxic content they can share, they are influenced by the non-toxic content they are exposed to, or they are disengaging with the platform. Overall, the proportion of shares that are toxic is positive, despite big reductions in the number of toxic posts shared. 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 4: Evidence on inelasticity in toxic sharing and seeking out behavior, by user type

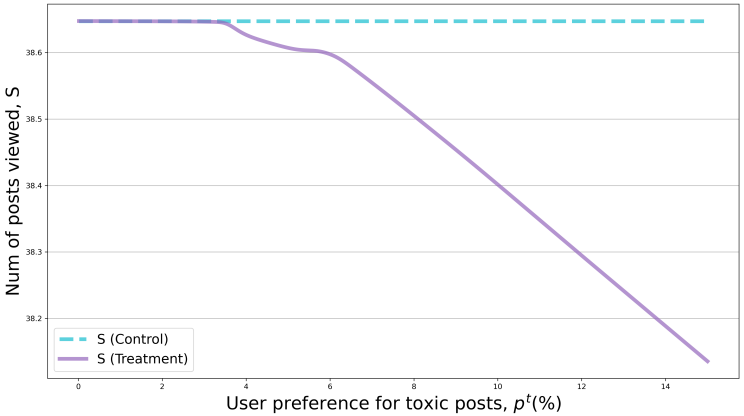


Notes: This figure provides complimentary evidence explaining the 'stickiness' in user behavior. User behavior is not malleable, because the ratio of toxic shares to toxic views is always less than 1 for all user types, across treatment and control groups. This means that user behavior is only partly influenced by the content they are exposed to, and this influence is quantified using the structural model. The first panel also suggests that toxic users are more malleable, while non-toxic users are mechanical, because there is no change in the proportion of toxic sharing to toxic viewing. The second panel shows that the treatment effect on the ratio of searches was positive and indistinguishable across user types. That is, toxic and non-toxic users were equally likely to seek out content, by using the 'text search' feature. 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.

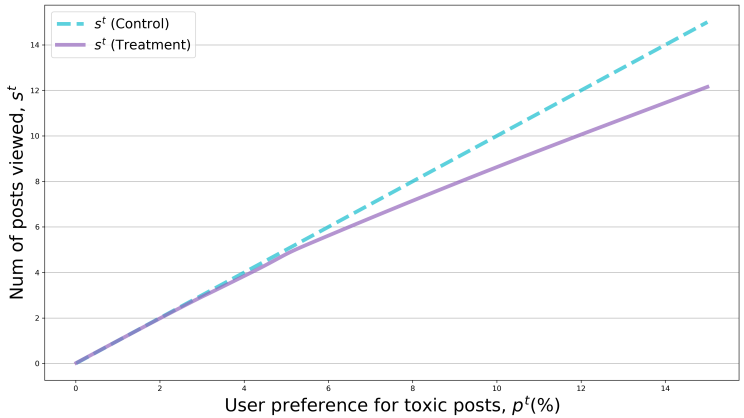
Figure 5: Model predictions, by user type or tastes for toxic content, p^t



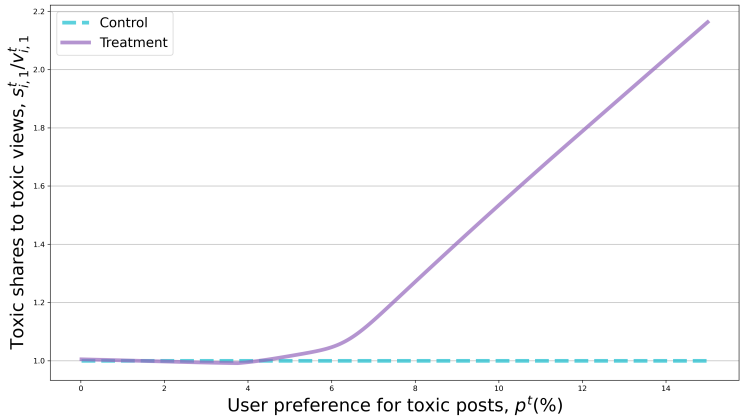
(a) Total number of posts viewed



(b) Total number of posts shared



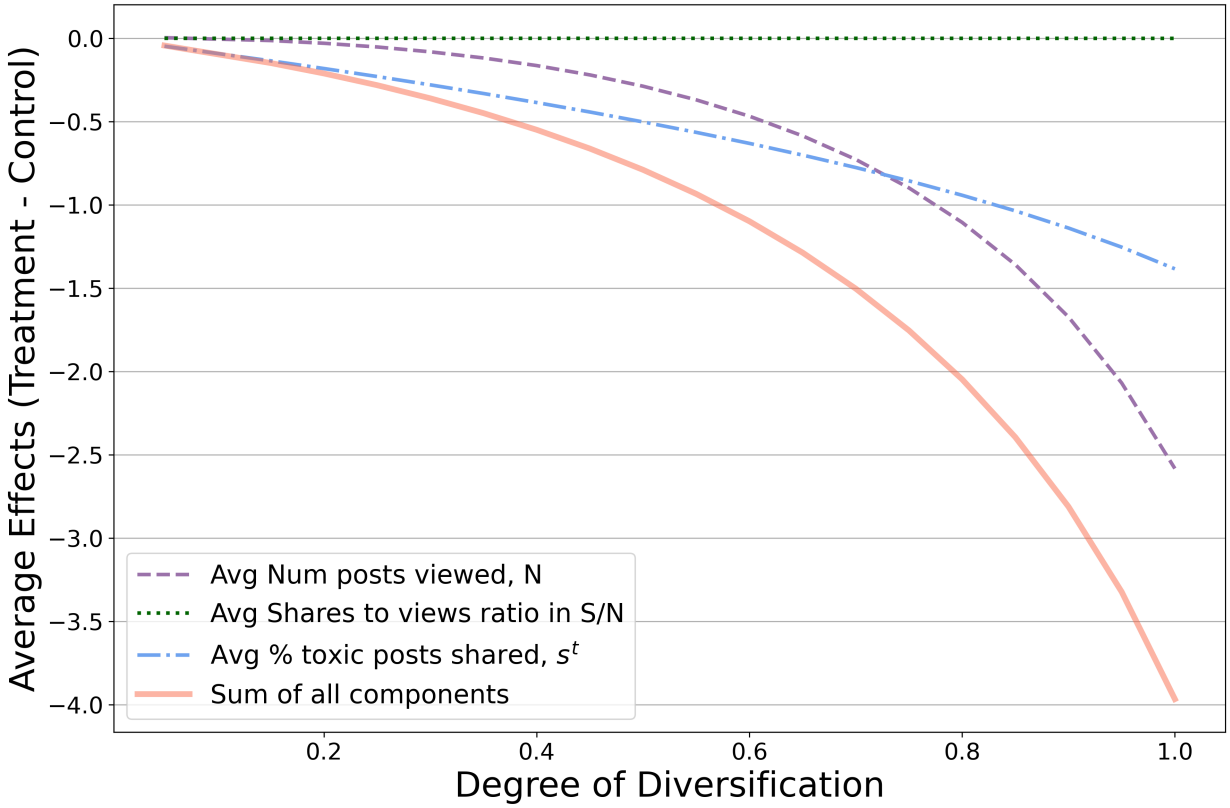
(c) Proportion of shares that are toxic



(d) Ratio of toxic shares to toxic views

Notes: This graph provides the model's predictions for key outcomes, for users with $p^t \geq \bar{q}^t$, where user type is defined by user tastes for toxic content, p^t. Panels (a) and (b) show that more toxic users (towards the extreme right in the p^t distribution) are expected to view and share smaller number of posts upon being treated. Panel (c) shows that the treatment effect on the proportion of shares that are toxic is expected to be negative for toxic users. This is due to two reasons: **(1)** the reduction in total usage of the platform is larger for toxic users, and **(2)** behavioral changes in the probability of sharing toxic content, due to reduced exposure to such content. Panel (d) predicts that the ratio of toxic shares to toxic views is monotonically increasing in p^t. These predictions are obtained using calibrated parameters from the structural model, by matching moment conditions for heterogenous users. Note that, for homogenous parameter values for users with different preferences, this model generates symmetric predictions for the number of posts viewed and shared. This is addressed later using different influence factors θ, for different user types.

Figure 6: Counterfactual policy predictions 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 a policy when $a = 60\%$, the decrease in the number of toxic posts is driven by the decrease in the probability of being assigned toxic content for toxic users. This is ideal for a policymaker who wants to reduce the number of toxic posts viewed and shared, without affecting the overall engagement of the platform. However, as the degree of randomization increases to 80%, decrease in engagement by toxic users 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 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 binary variable taking value 1, when user i was assigned to the treatment group. The table shows a randomly selected set of coefficients. Weeks correspond to the date on which a user created her account. None of the observable characteristics are correlated with treatment assignment. I cannot reject null hypothesis of joint insignificance, with an F-statistic of 0.984 and p-value of 0.506. The regression was estimated at the user level. Robust standard errors in parantheses.

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 Posts Shared
Treatment Effect	-0.641**	0.120**
	(0.033)	(0.038)
Control Mean	7.416**	1.547**
	(0.018)	(0.018)
N	231814	

Notes: This table shows that the treatment effect on the number of posts viewed and shared, as well as the number of toxic posts viewed and shared, in one month, is negative and statistically significant. Each cell reports estimates of the regression coefficient from a linear regression of the outcome aggregated at the user level, over days in the first month of the intervention period (February 10 to March 10, 2023). Robust standard errors are in parentheses. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

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)	(0.0580)	
Proportion of Toxic Posts Viewed (Baseline)			-0.183**
			(0.0652)
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 behavior, according to the perceived social norms. θ is, therefore, the influence effect of one month's exposure to non-personalized feeds. This is modelled as the extent to which users share content to signal their conformity with the behavior of other users in their network, to derive benefits of public recognition. Column (1) shows relevance of the instrument, i.e. the differences between probability of viewing toxic and non-toxic content, computed using only the first half of posts viewed by user at baseline, when they were arranged in a random order ($half1$). This instrument is used to correct the measurement error, on account of treated users randomly sampling toxic posts to view from their feeds. The independent variable in Column (1) is the difference between proportion of toxic and non-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 results of a 2SLS regression of the difference between baseline and intervention period in differences between probability of *sharing* toxic and non-toxic content. Here, the independent variable is $half1$, which was 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. Estimated slope coefficient estimates γ_1 is always negative and statistically significant. Estimated θ is therefore, positive, and estimated to be 0.16, according to the preferred specification in Column (2). Baseline period is December, 2022 and intervention period data spans from February 10, 2023 and March 10, 2023. Robust standard errors are in parentheses. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

A 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.

A.1 Social Media and Indian Politics

The harms of social media have garnered significant attention in the US, but are arguably more severe in India. This is because as more Indian 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 (Mukherjee, 2020).

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.[35]

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. In agreeing with the government, social media platforms may be biased against opposition parties and pressure groups. 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[36].

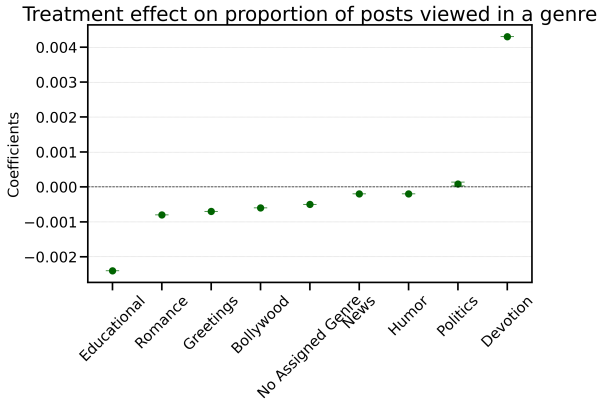
[35]See `https://thewire.in/tech/meta-approved-ai-manipulated-political-ads-during-indias-election-report`

[36]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.

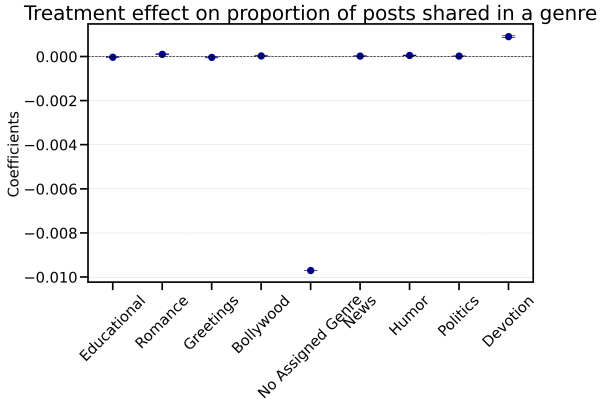
A.2 SM: 'Indian TikTok'

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.[37] While the platform is home to organic content creators, various politicians, and Bollywood celebrities also sometimes interact with their follower base on this platform.

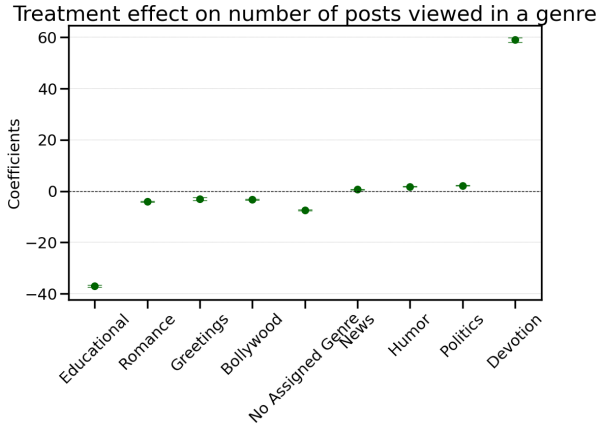
Figure A.1: Treatment effects on viewing and sharing content from various genres



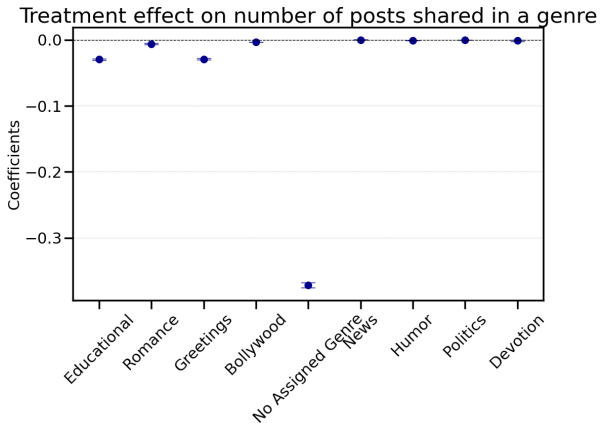
(a) F-stat: 6122.932, p-value: 0.000



(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.

[37] 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 A.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 A.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 the 'Groups' they join.

Figure A.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.'

B 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.

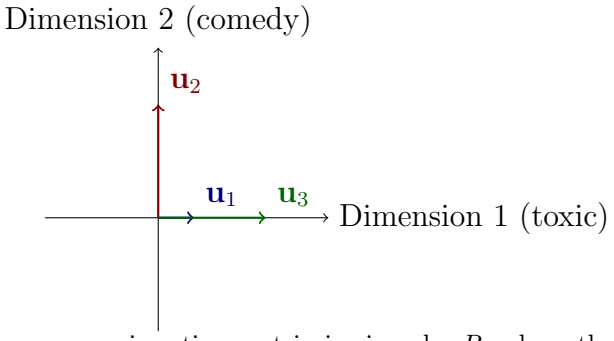
Figure B.1: An example of SVD decomposition into two-dimensional user embeddings U, eigenvalues Σ, and movie embeddings V^T

(a) Matrix representation of vector embeddings

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

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

(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. Then, the user embeddings would show how much each user likes toxic and comedy movies. In this example, the first user is more interested in toxic movies, while the second user is more interested in comedy movies.

B.1 Illustration: Control Algorithm

Consider an example with three users and two movies in Figure B.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 for each user. 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 (Hastie et al., 2015). 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.[38]

B.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 randomly select 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 B.2: Matrix representation of vector embeddings, for treated and control users

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

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

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.

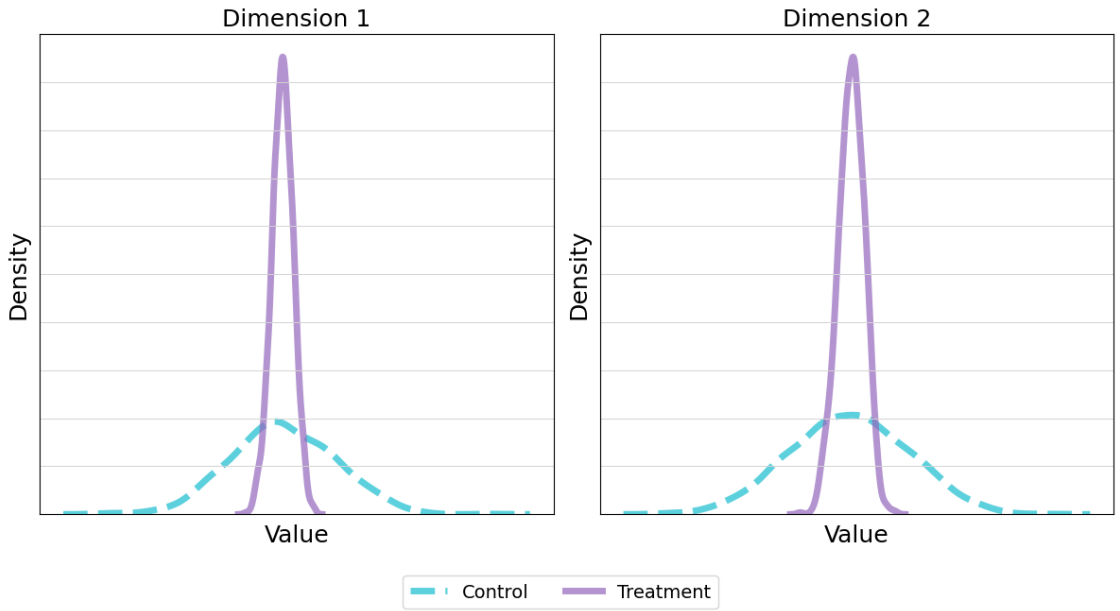
This example makes another subtle point. The embeddings generated for each treated user are equal to the average of the embeddings for the control group users. Therefore, there

[38] See (Ludwig and Mullainathan, 2024) for a recent and highly innovative use of contextual embeddings in a labor economics application.

is not enough variation in the embedding assignment within the treatment group, as the treatment embeddings are concentrated around the mean embedding value, by application of the Central Limit Theorem (CLT). This is depicted in Figure B.3, for the simulated (two-dimensional) recommendation algorithm. This necessitates the need for a structural model to identify the effect of exposure on engagement.

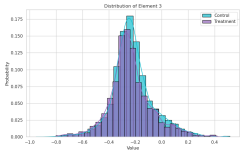
Figure B.4 shows the distribution of a randomly selected set of user embeddings. As expected, the treatment group embeddings are more concentrated around the mean than the control group embeddings. Further, the CLT predicts that the treatment group embeddings follow a normal distribution, with a variance smaller than the control group embeddings. The dimensions of these embedding vectors could not be interpreted in any human-intelligible terms.

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

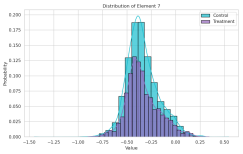


Notes: This graph shows that the two dimensions (components) of the embedding vector follow a Gaussian distribution, where the embeddings were simulated using a simple SVD algorithm and a matrix of engagement in the control group. An embedding is a representation of complex data in a lower-dimensional 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 (CLT).

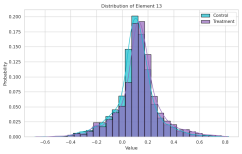
Figure B.4: Treatment effect on embeddings across various Dimensions



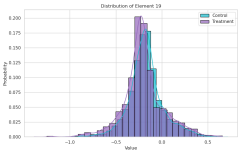
(a) 3



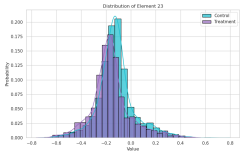
(b) 7



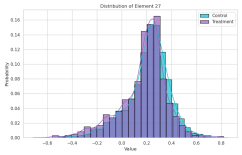
(c) 13



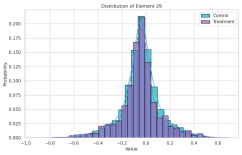
(d) 19



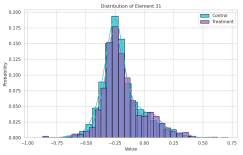
(e) 23



(f) 27



(g) 29



(h) 31

Notes: This figure shows the empirical distributions of randomly selected dimensions of user embeddings in the treatment and control groups. On average, the user embeddings for the treatment group were more concentrated around the mean than the control group, as predicted by the simulated embeddings.

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.[39] However, Table B.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.

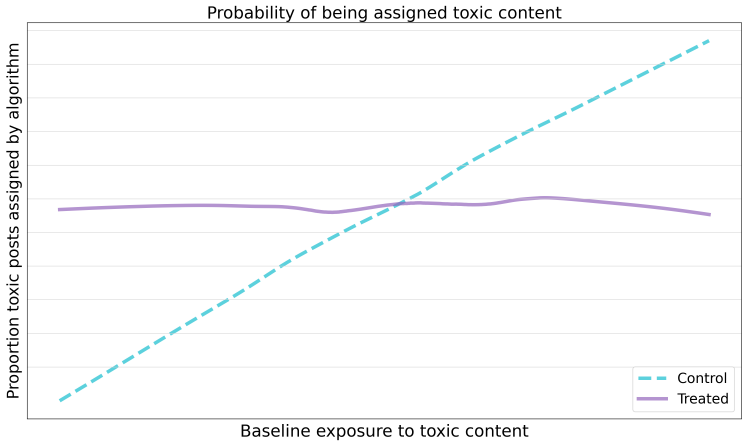
Table B.1: Popularity of posts viewed by users in the treatment group

	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	

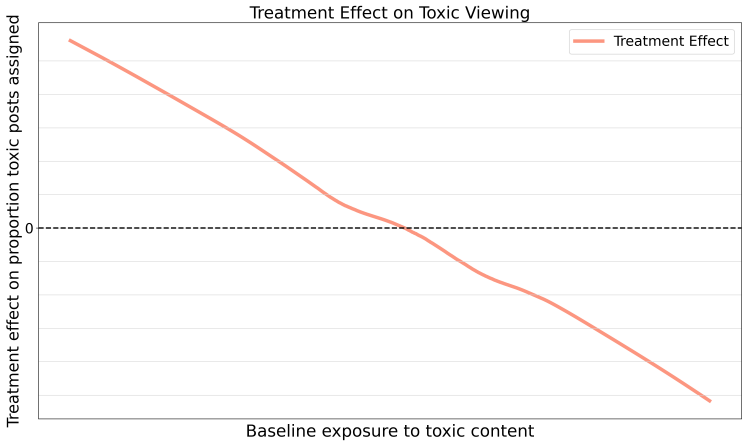
Notes: This Table shows that, contrary to expectations, the treatment group was exposed to less popular posts than the control group. It is possible that in bringing the treatment group's preference weights closer to those of an average user, the intervention recommended posts are more appealing to the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

[39]This follows from the logic of the main predictions of the median voter theorem.

Figure B.5: 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. Details of the simulated personalization algorithm, and the intervention's random algorithm, are in Appendix **??**.

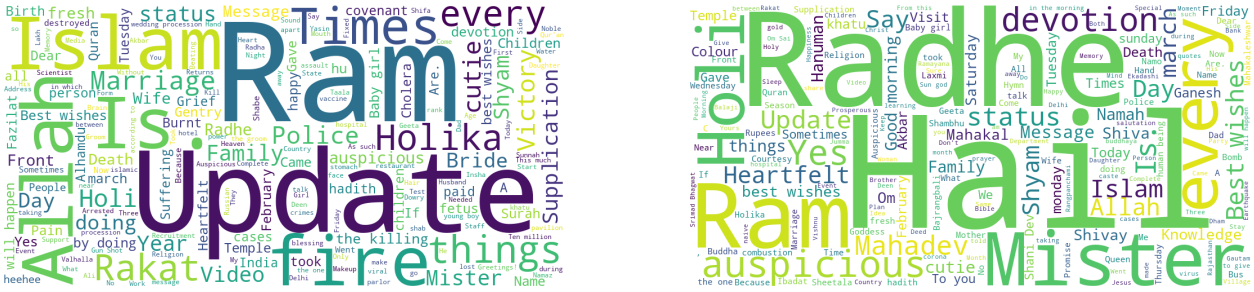
C Text Analysis

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.

C.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 C.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 consumer or post content.

Figure C.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 C.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.[40] 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).

[40]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 (Handlan, 2020).

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.

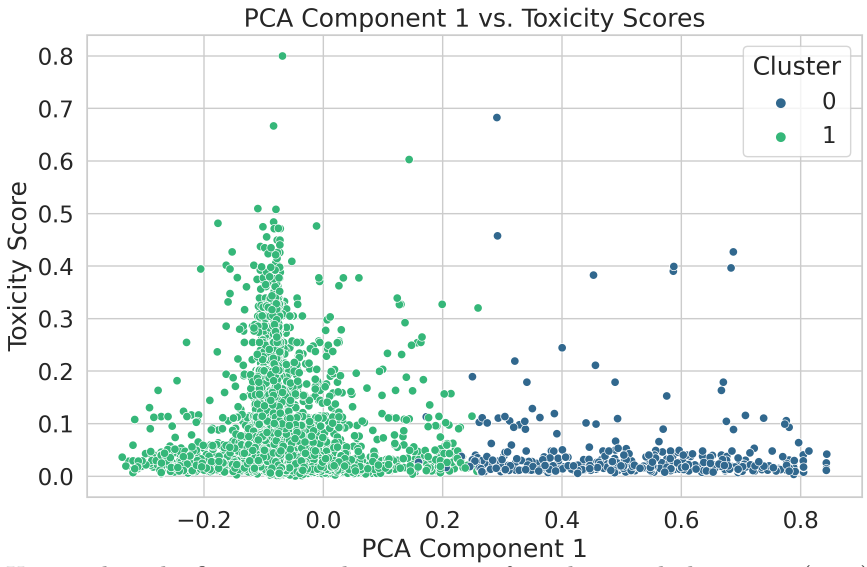
C.2 Toxicity Algorithm

In keeping with the literature on social media harms, I use the Perspective API to classify posts as toxic or not (Aridor et al., 2024). 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 C.1.

Figure C.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.

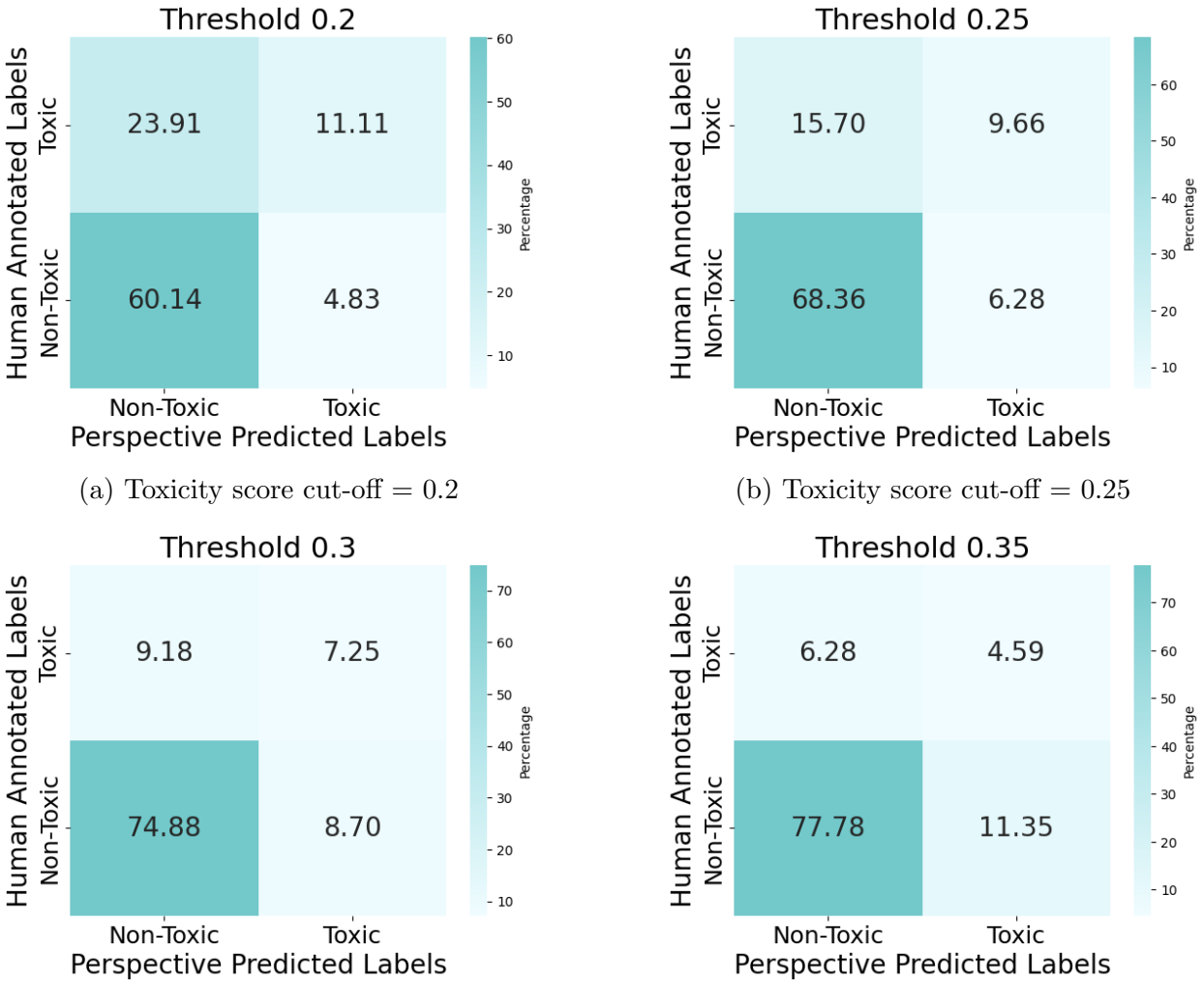
In Figure C.2, I plot and cluster the first principal component of the vectorized TF-IDF word representation for each document, against the corresponding toxicity scores for that post (Gentzkow et al., 2019; Ash and Hansen, 2023). I find that the separation between the word clusters corresponds to the 0.2 cut-off in the toxicity measure. 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 C.3 show that the 0.2 cut-off has the best performance in terms of correctly classifying toxic posts. This criterion is important because toxicity is a rare outcome and can, therefore, make automatic detection difficult (Banerjee et al., 2023).

Figure C.2: Validating toxicity score threshold to construct binary outcome



Notes: The X-axis plots the first principal component of words in each document (post), obtained using PCA on the TF-IDF matrix of words. Y-axis corresponds to the toxicity score for the corresponding documents (posts), computed using the Perspective API. A k-means clustering algorithm is used to cluster the posts based on the first principal component. Toxicity scores for most posts in the first (blue) cluster of the first principal component rarely exceed 0.2. Therefore, 0.2 is chosen as the appropriate threshold for the binary outcome variable, indicating whether a post is toxic or not.

Figure C.3: 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.

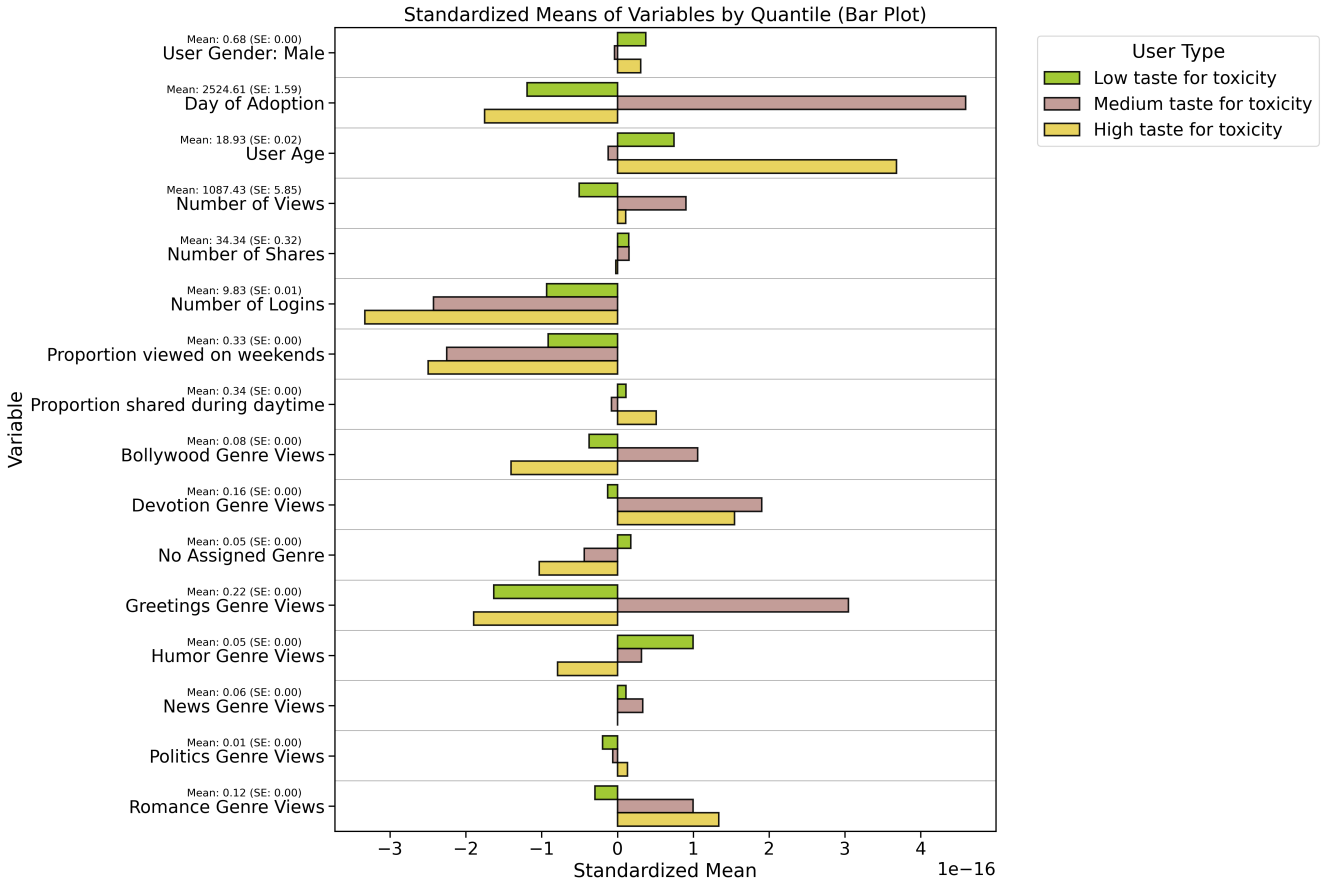
Table C.1: Examples of text data (English translations) with toxicity scores

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.

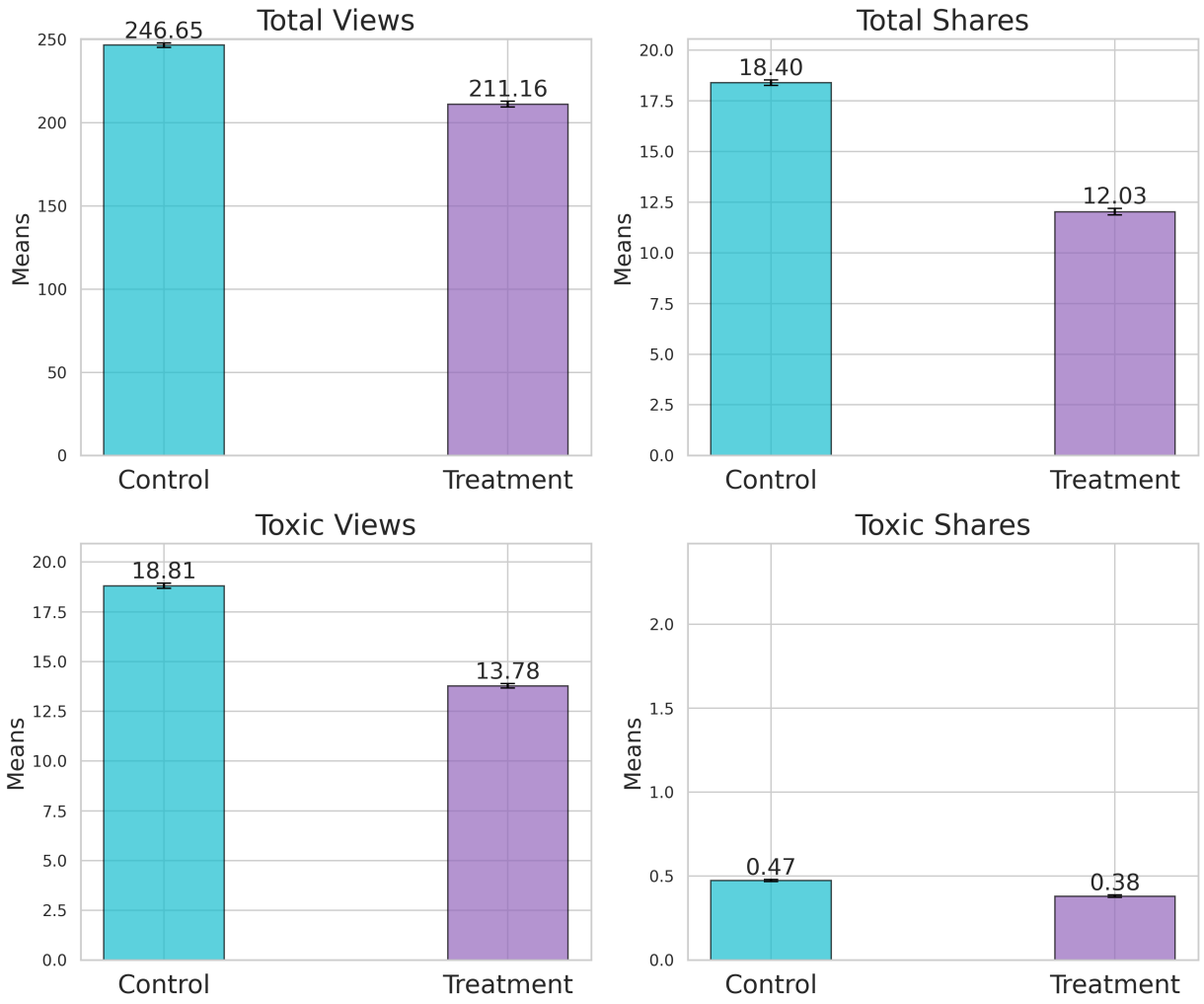
D Supplementary Figures

Figure D.1: Key user attributes at baseline



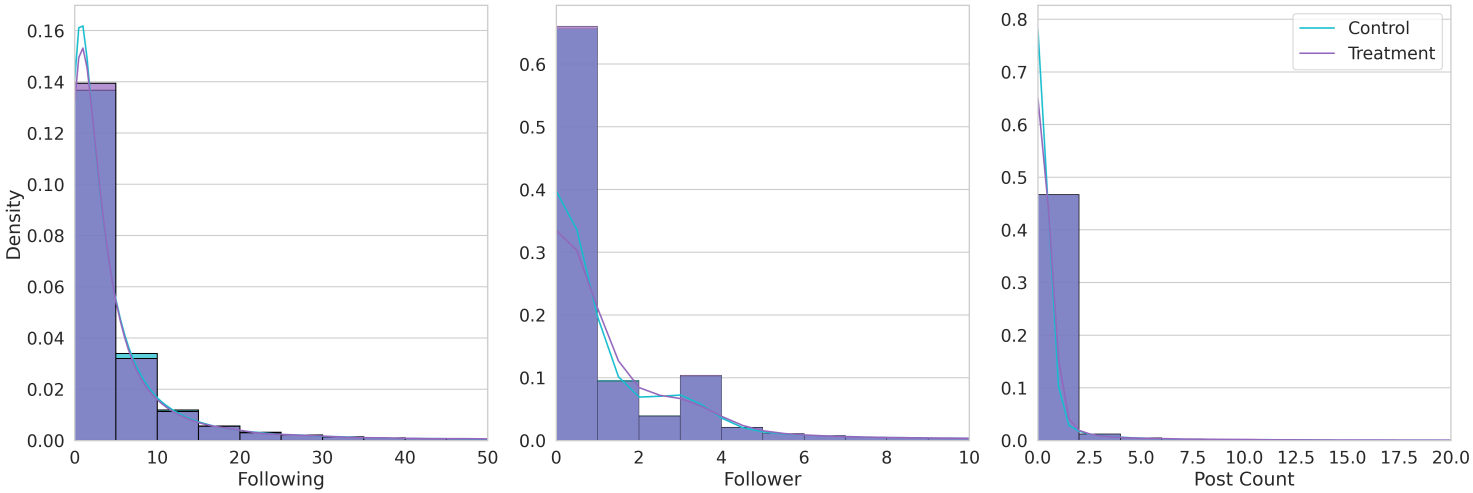
Notes: This Figure shows the baseline attributes of users in the experimental sample, distributed across user type. User type is defined by the proportion of toxic posts viewed at baseline, and users are allocated to the quantile in which they fall. The bar charts are constructed after standardizing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

Figure D.2: Comparison of means across treatment and control, for key outcomes



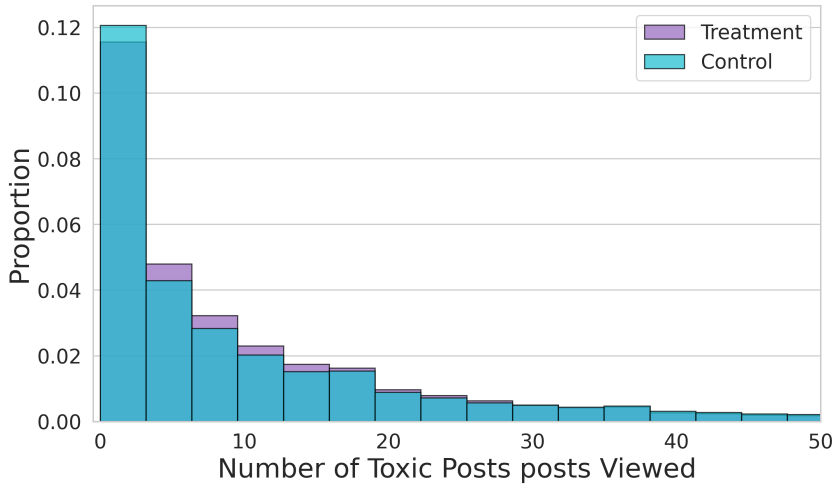
Notes: This Figure shows the trade-off between user engagement and the propagation of harmful content on social media. During the first month of the intervention, treated users were, on average, exposed to less toxic content, but were also less active on the platform. This highlights the costs (in terms 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.3: Distinct features of SM's interface

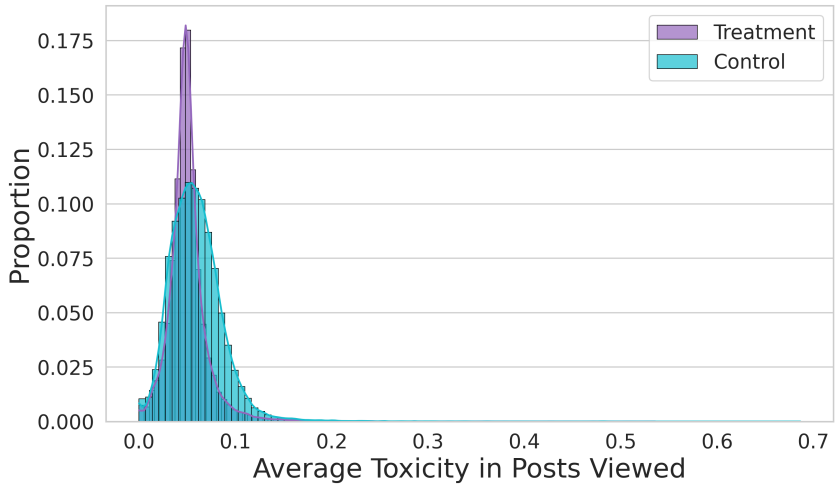


Notes: This figure shows that SM's user interface is distinct from other social media platforms. In particular, despite having an option to follow other users, the platform is content-based, and users interact with content rather than with other users. This is in contrast to platforms like X (formerly, 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.4: Distribution of exposure to toxic content during intervention period



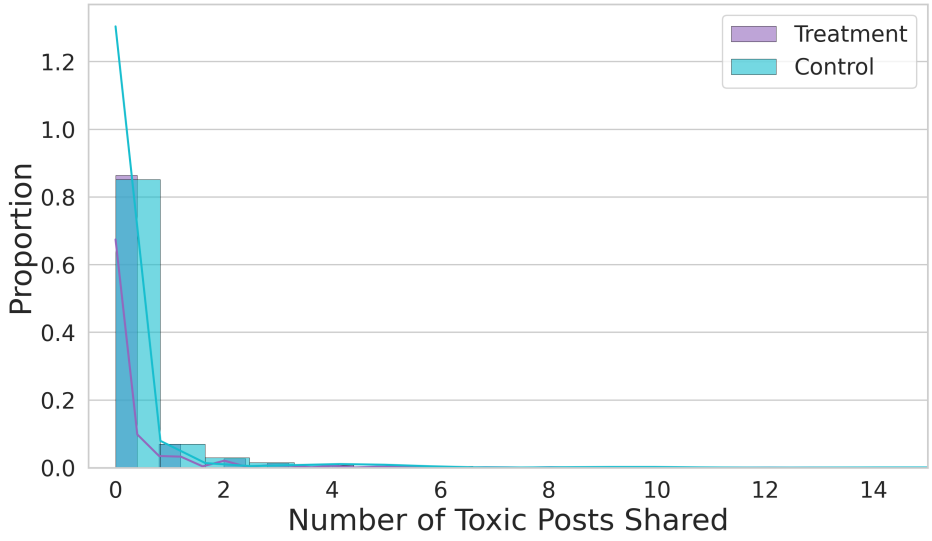
(a) Number of toxic posts viewed (with binary indicator for post toxicity)



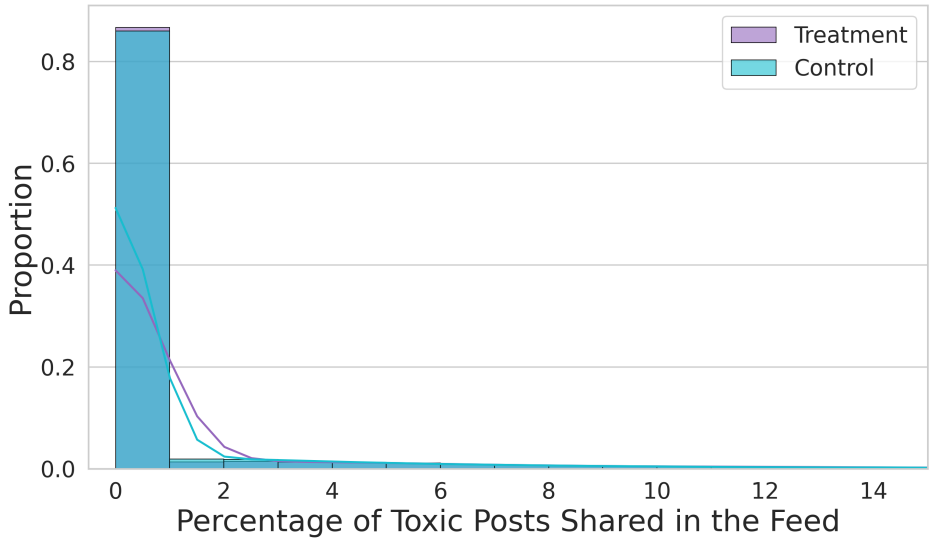
(b) Average toxicity scores on posts viewed (with continuous toxicity scores)

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.5: Distribution of engagement with toxic content during intervention period



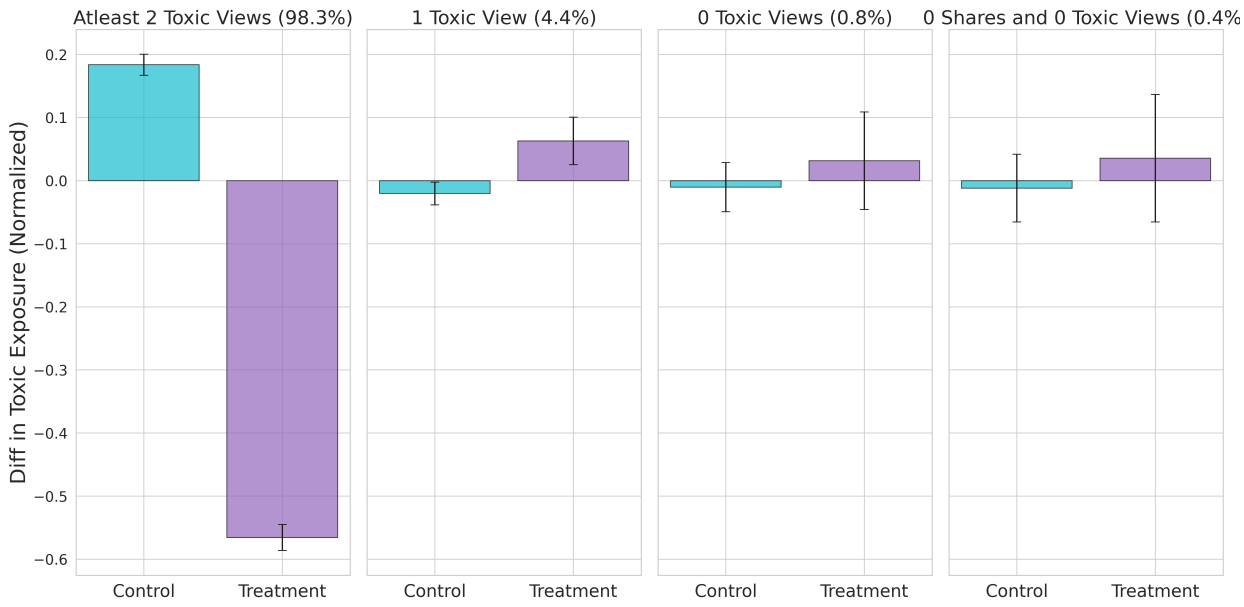
(a) Number of toxic posts shared



(b) Percentage of posts shared that are toxic

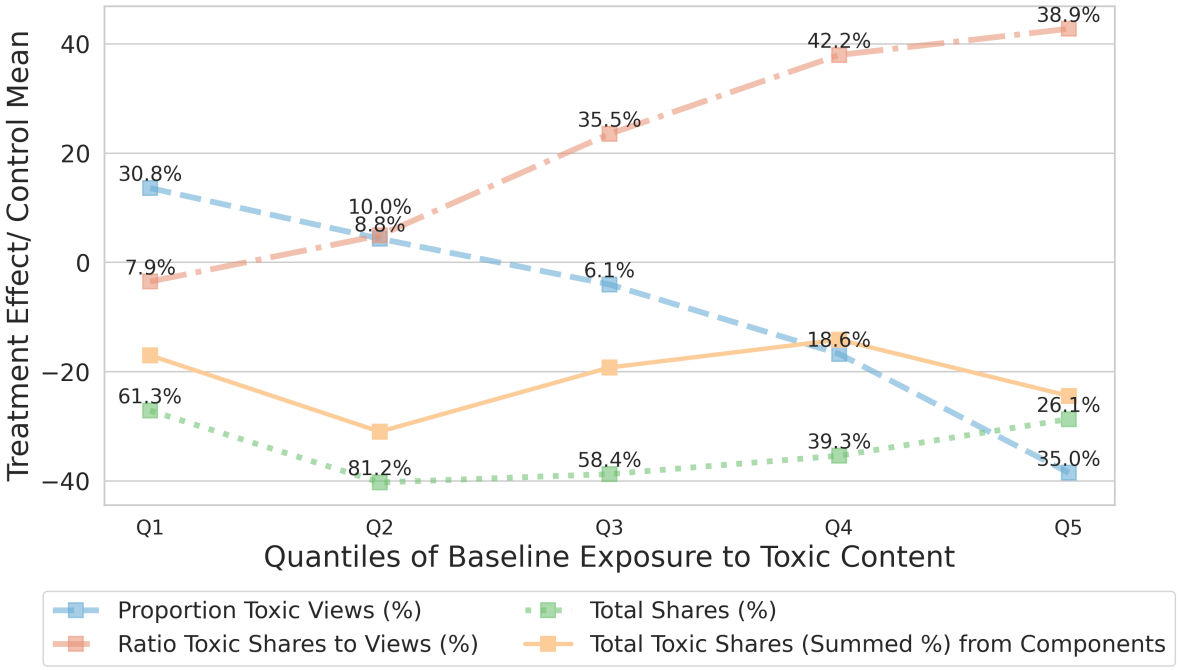
Notes: This Figure plots the raw data on toxic shares for treated and control users, and shows that the distribution toxic posts shared by control users first order stochastically dominates the distribution for treated users. Panel (a) provides the number of toxic posts shared, where a toxic share is 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.6: Difference between toxic views at intervention period and baseline, by user engagement at baseline



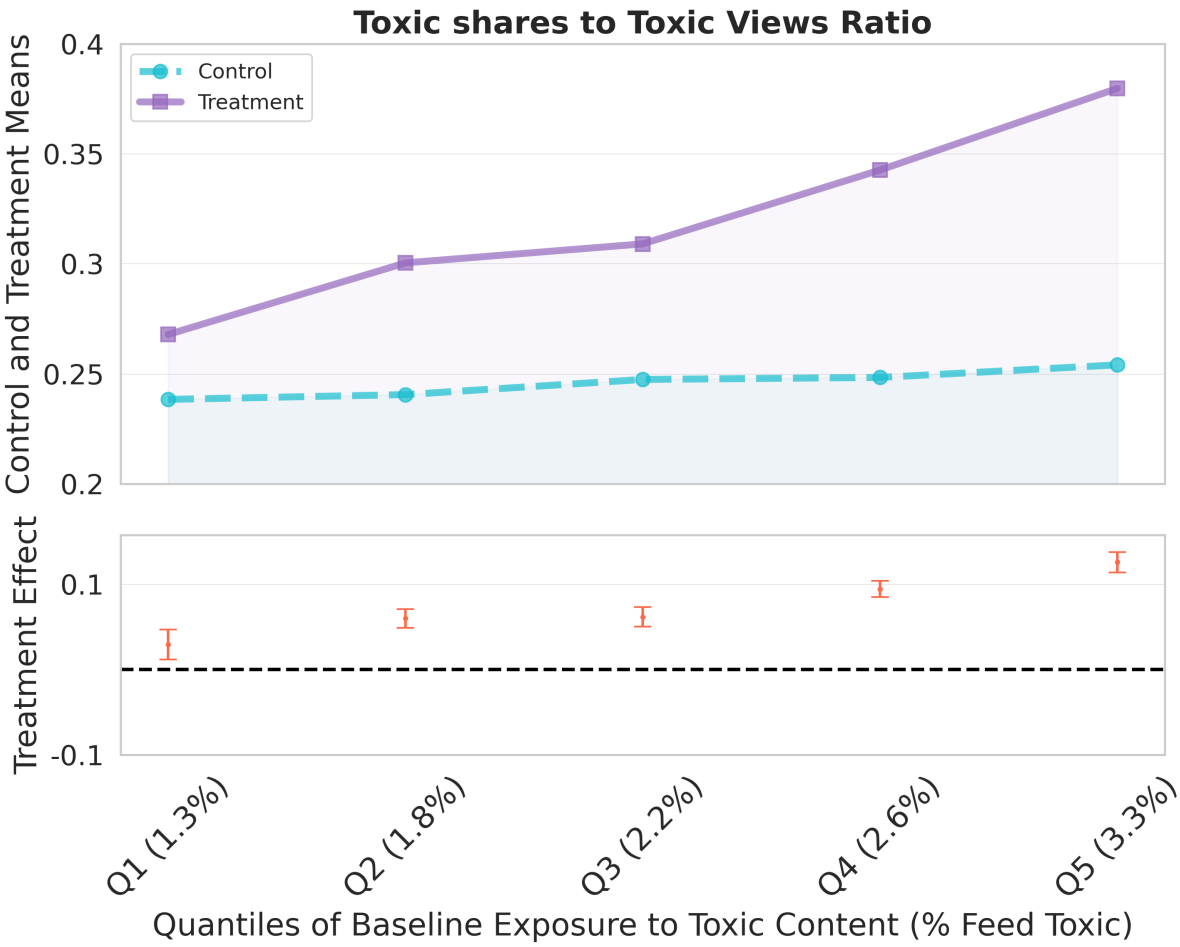
Notes: This Figure plots the raw data showing difference between toxic views during the intervention period and baseline, for users with different levels of toxic engagement at baseline. The averages have been normalized to have a mean of 0 and a standard deviation of 1. Percentage of users of a type (by engagement at baseline) in the experimental sample is given in parentheses. For example, 98.3% of users in the sample viewed at least two toxic posts at baseline. This Figure shows that these users saw fewer toxic posts during the intervention period. Users who saw exactly one toxic post at baseline, form about 4.4% of the experimental sample, and saw a larger number of toxic posts if they were treated.

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



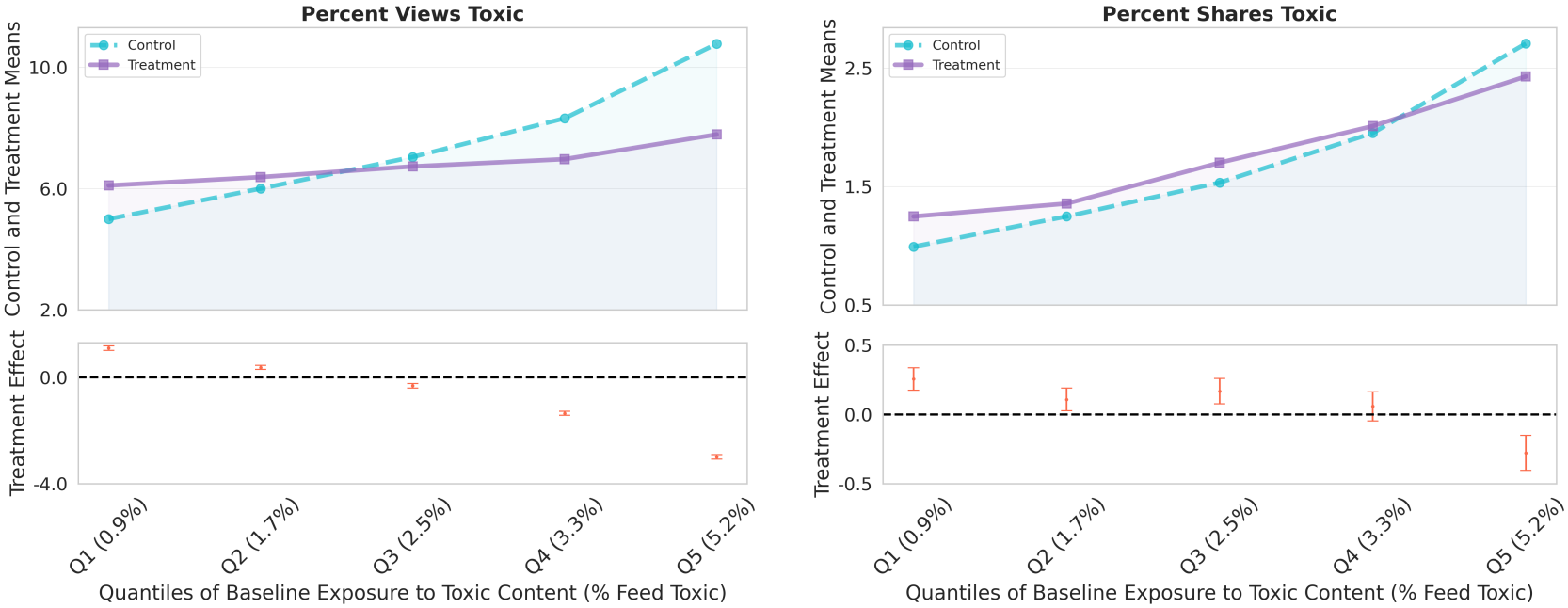
Notes: This Figure shows replicates the main result, that the ratio of toxic shares to toxic views is increasing in user's baseline toxicity, even when toxicity is measured by averaging over the continuous toxicity scores of posts viewed or shared. On average, the ratio of toxic shares to toxic views 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



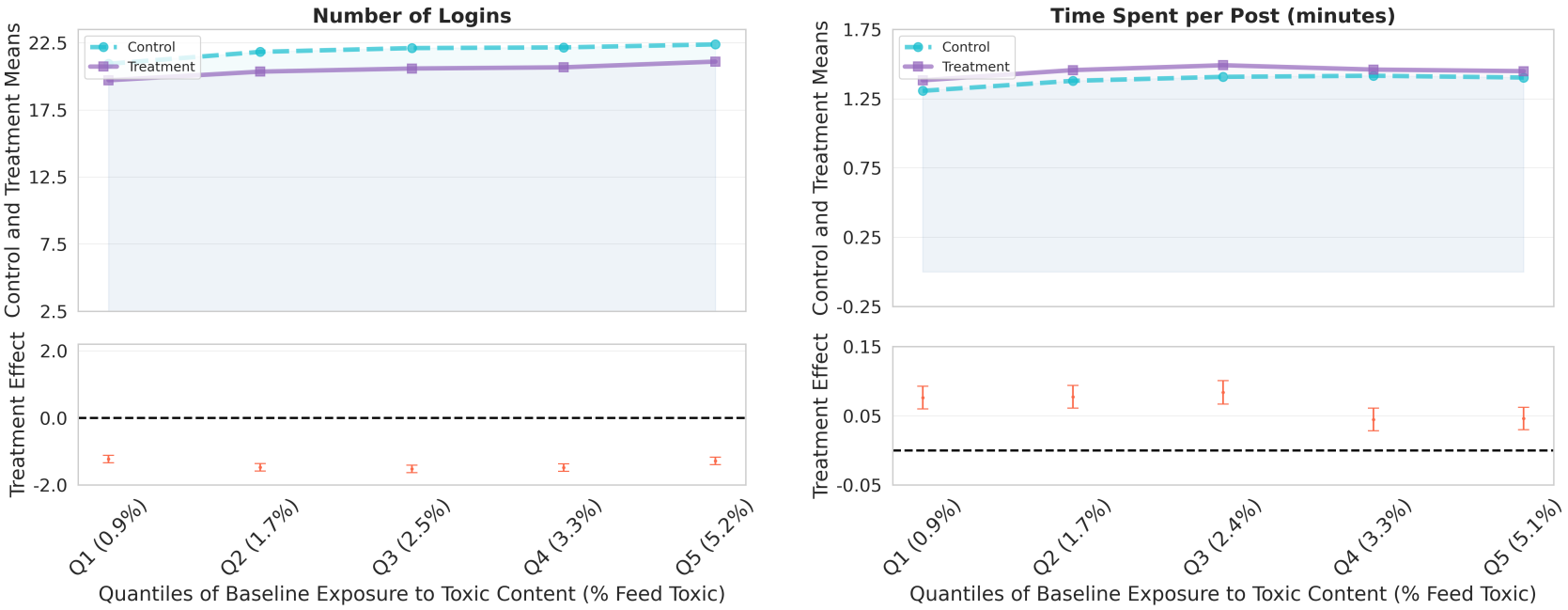
Notes: This Figure shows replicates the main result, that the ratio of toxic shares to toxic views is increasing in user's baseline toxicity, even when toxicity is measured by averaging over the continuous toxicity scores of posts viewed or shared. On average, the ratio of toxic shares to toxic views 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)



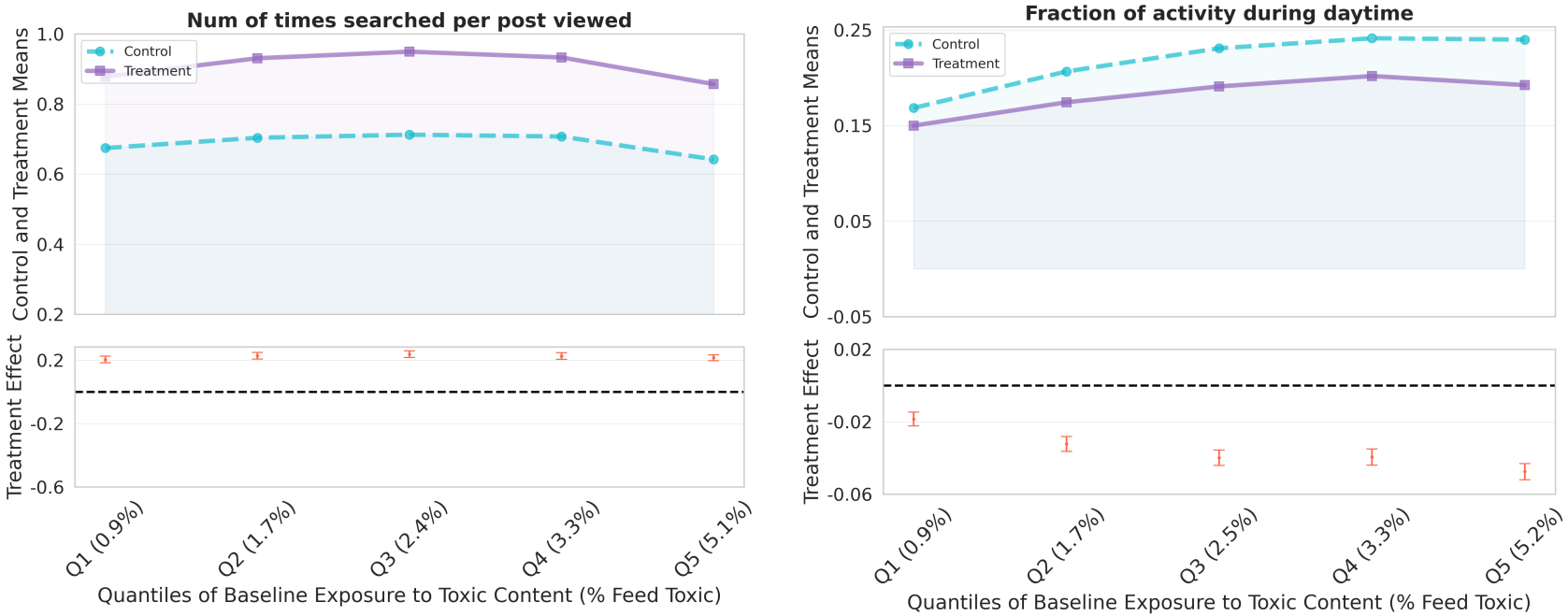
Notes: This figure shows that the treatment effect, on the *proportion* of posts shared that are toxic, is *non-negative* for all users except those in Q5 (with the highest exposure to toxic content at baseline). This is true, even in the cases of Q3 to Q5, where user type is toxic enough that the treatment 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 platform activity, by user type



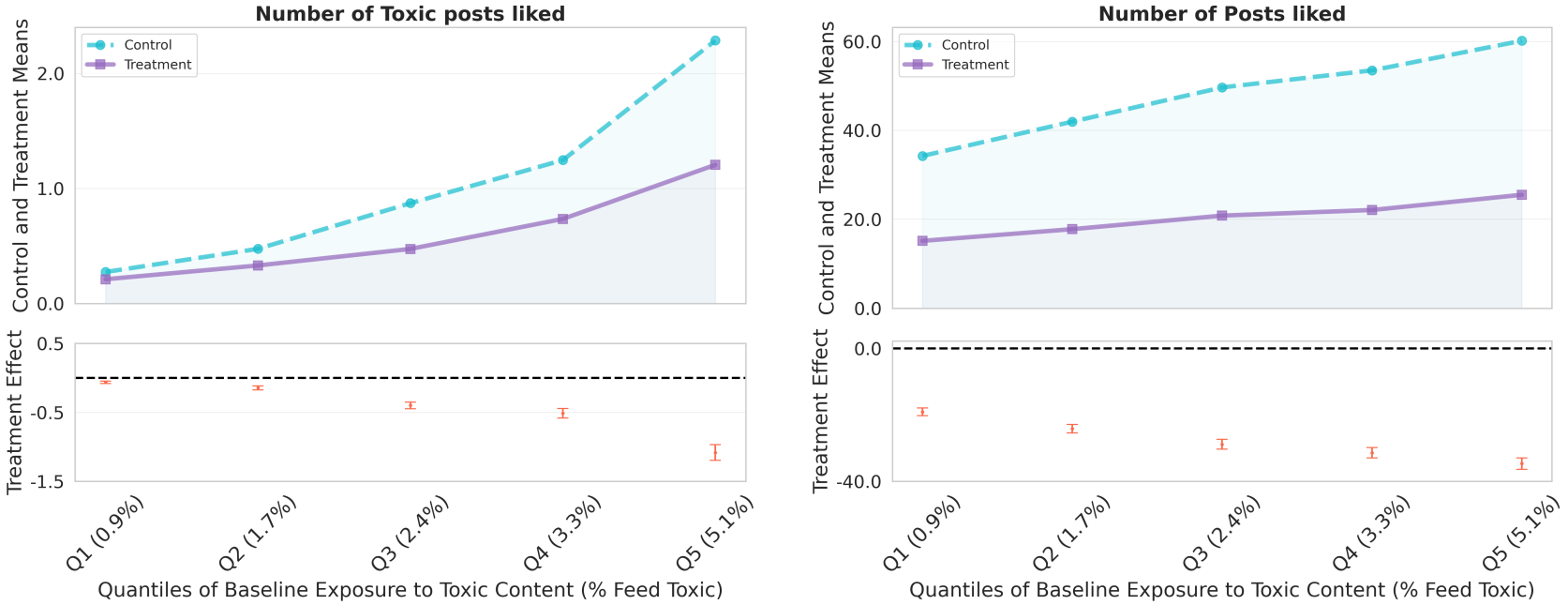
Notes: This figure shows that the treatment effect, on the overall activity of users on the platform is negative for all users. There is a decrease in the number of times a user logged in. The effects on logins are indistinguishable across user types. There is an increase in time spent per post, but 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.11: Auxiliary evidence of users seeking out preferred content



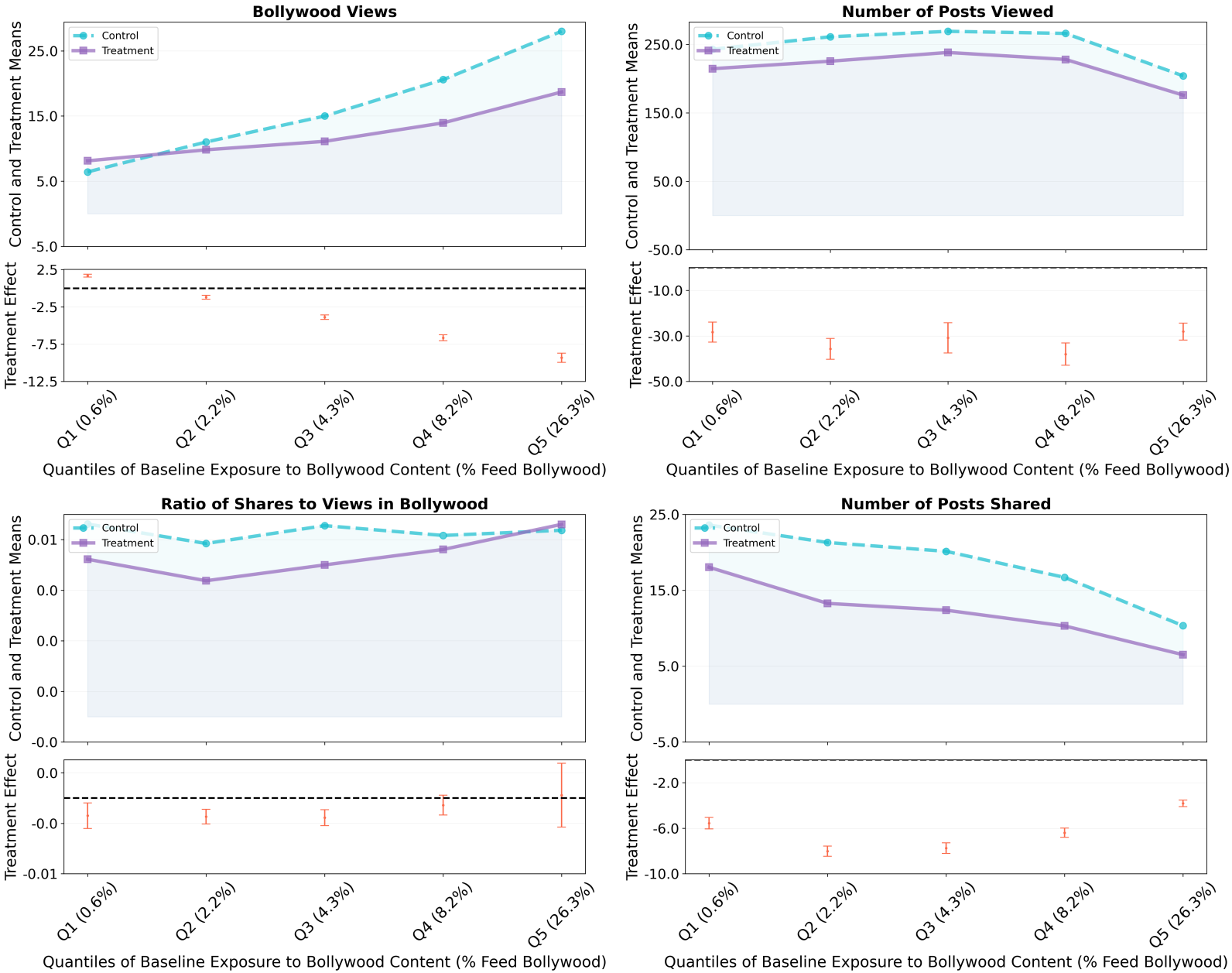
Notes: This figure shows that the intervention changed the quality of user engagement on the platform. This is because treated users were more likely to use the platform during the weekend, or in the night. Users spent less on the platform during their working hours. This has significant implications 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.12: Experimential effects on Liking behavior, by user type



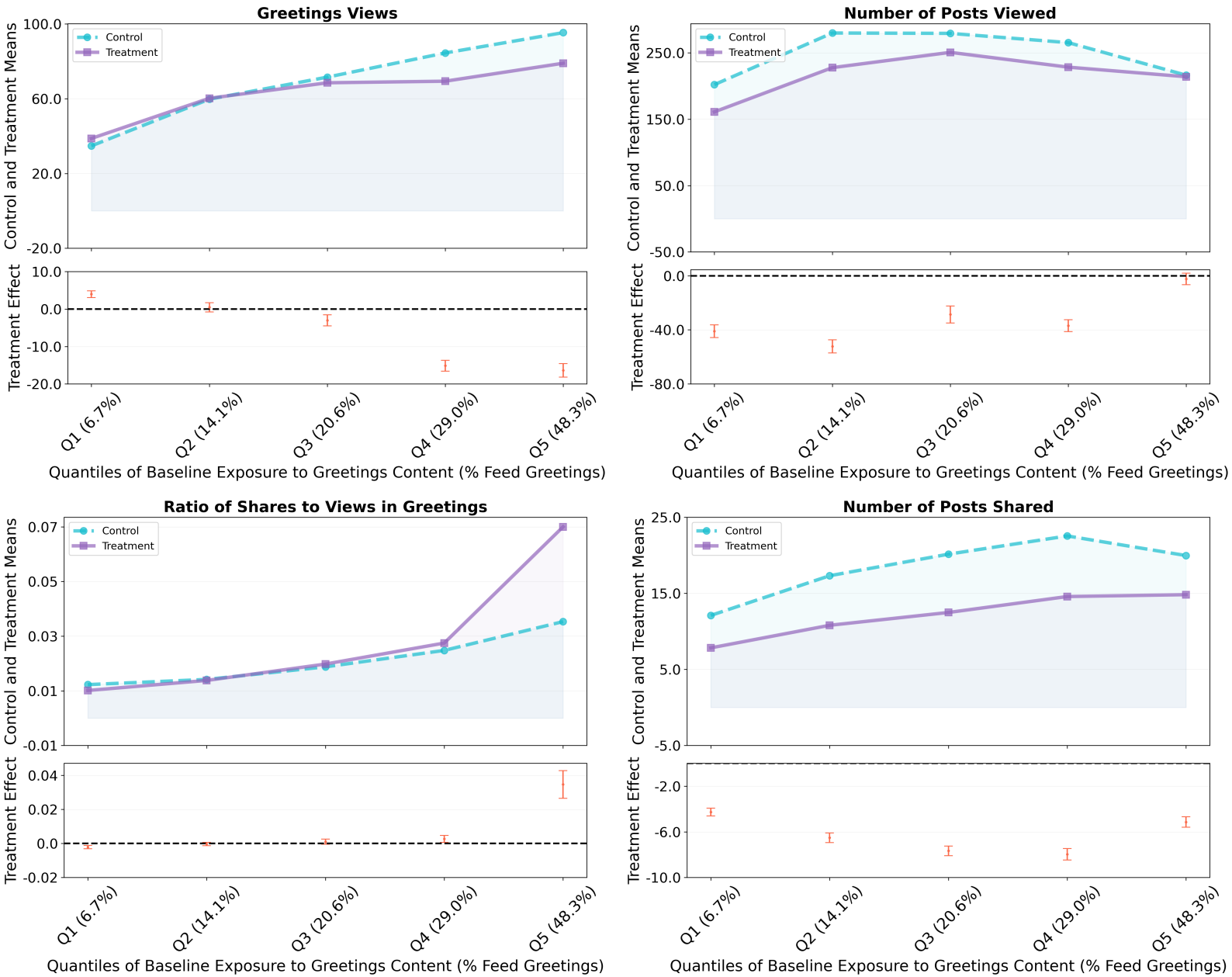
Notes: This figure shows that the treatment effect, on the *proportion* of posts shared that are toxic, is *non-negative* for all users except those in Q5 (with the highest exposure to toxic content at baseline). This is true, even in the cases of Q3 to Q5, where user type is toxic enough that the treatment 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.13: Treatment intensity with respect to Bollywood content



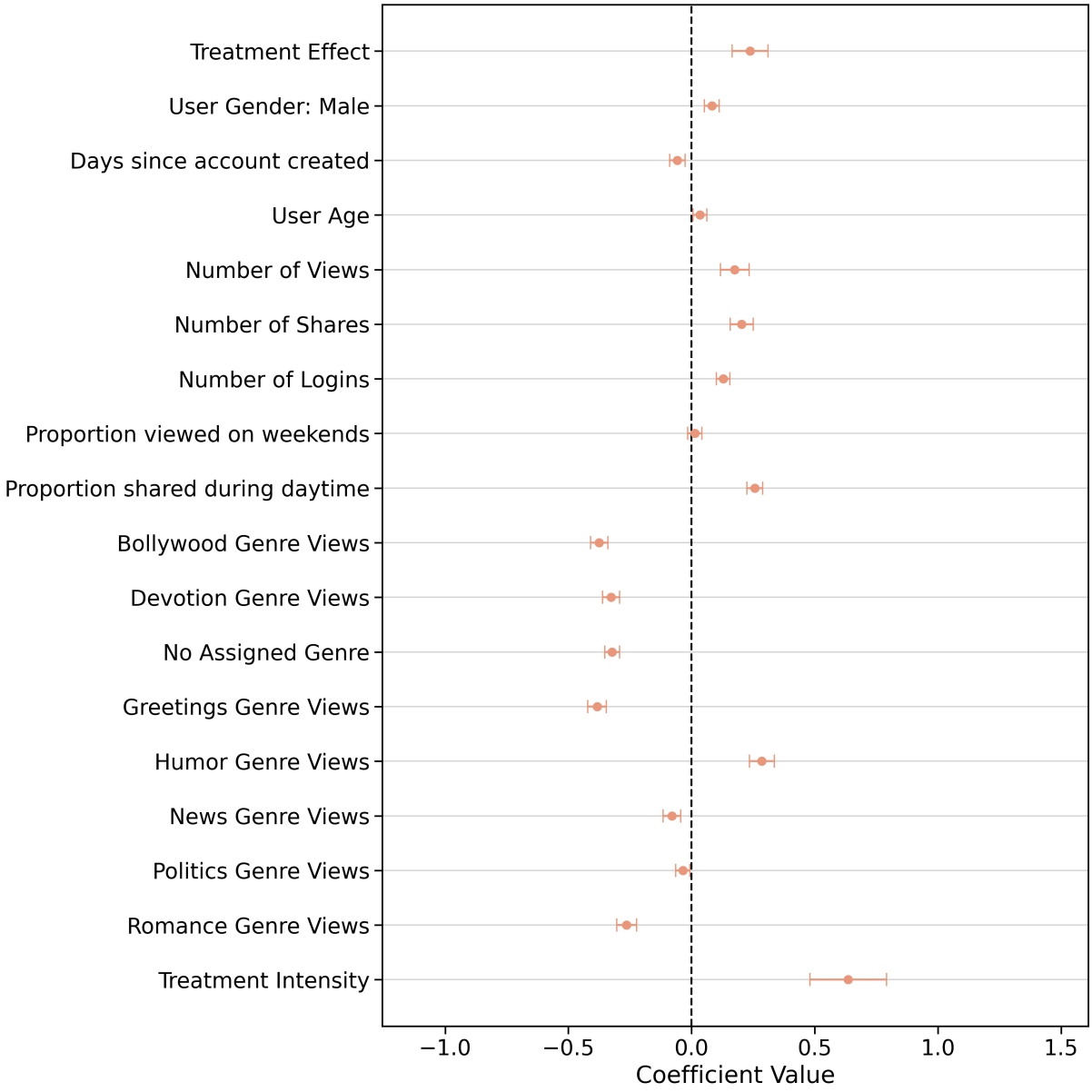
Notes: This Figure shows that the treatment effects on the number of Bollywood related posts mimics the treatment intensity with respect to toxic content. However, the treatment effects on the total number of posts viewed (of any type) do not follow the same pattern, according to user type defined in 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.14: Treatment intensity with respect to Greetings content



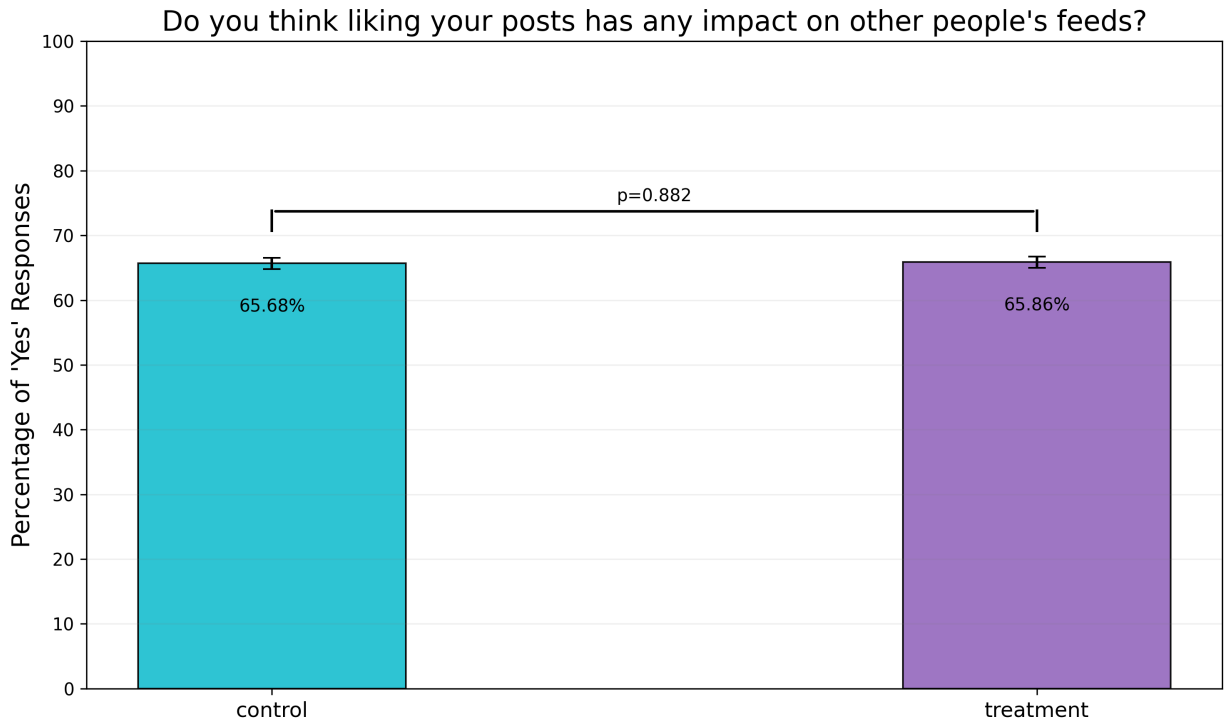
Notes: This Figure shows that the treatment effects on the number of Greetings related posts mimics the treatment intensity with respect to toxic content. The effects on the number of posts viewed in this genre, and the ratio of shares to views in the Greetings category, follow patterns similar to those 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.15: Suggested mechanisms driving engagement with toxic content



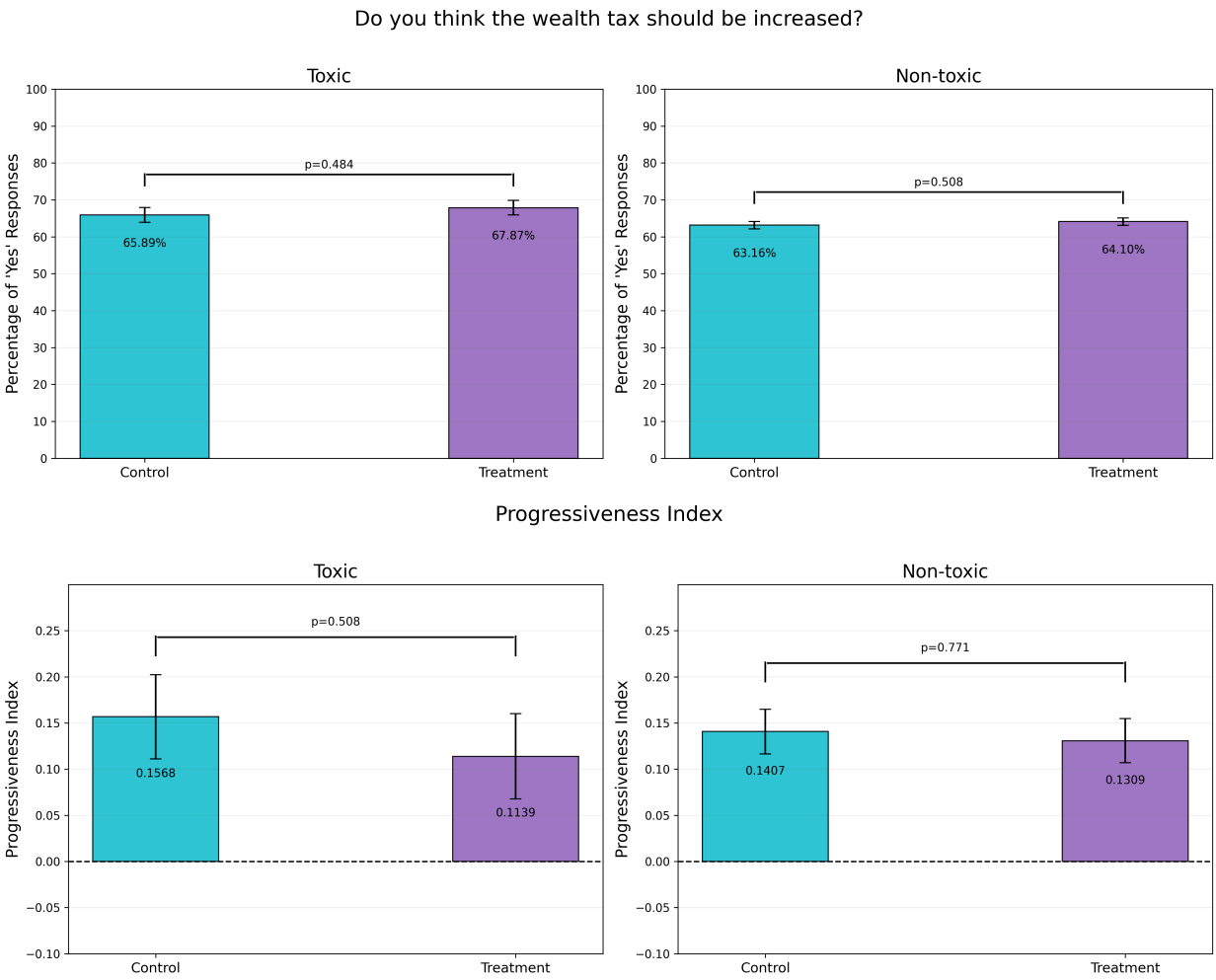
Notes: This Figure shows suggestive evidence on the mechanisms driving the treatment effects, by regressing the main outcome variable (proportion of shares that are toxic), on treatment status, treatment intensity (proportion of views that are toxic), as well as baseline user characteristics. For all 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.16: Salience of personalization algorithms



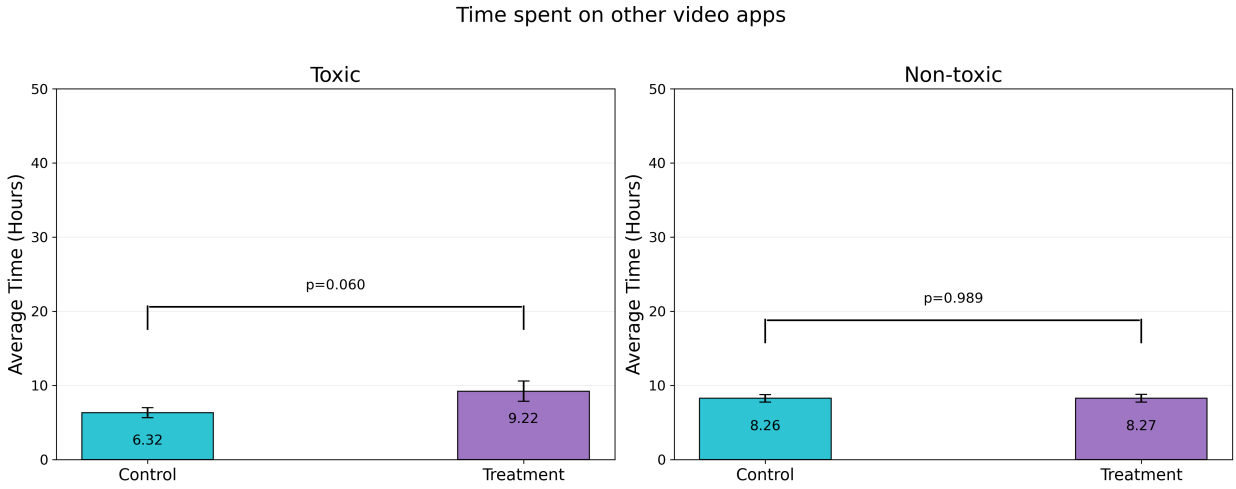
Notes: This Figure shows that the personalization algorithm is salient to users. A subset of users in the experimental sample were randomly selected for a follow-up survey ($N = 8,387$), and asked whether they thought their likes and shares changed the content in other users' feeds. More than 65% of 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.17: Preferences over redistribution, by user type



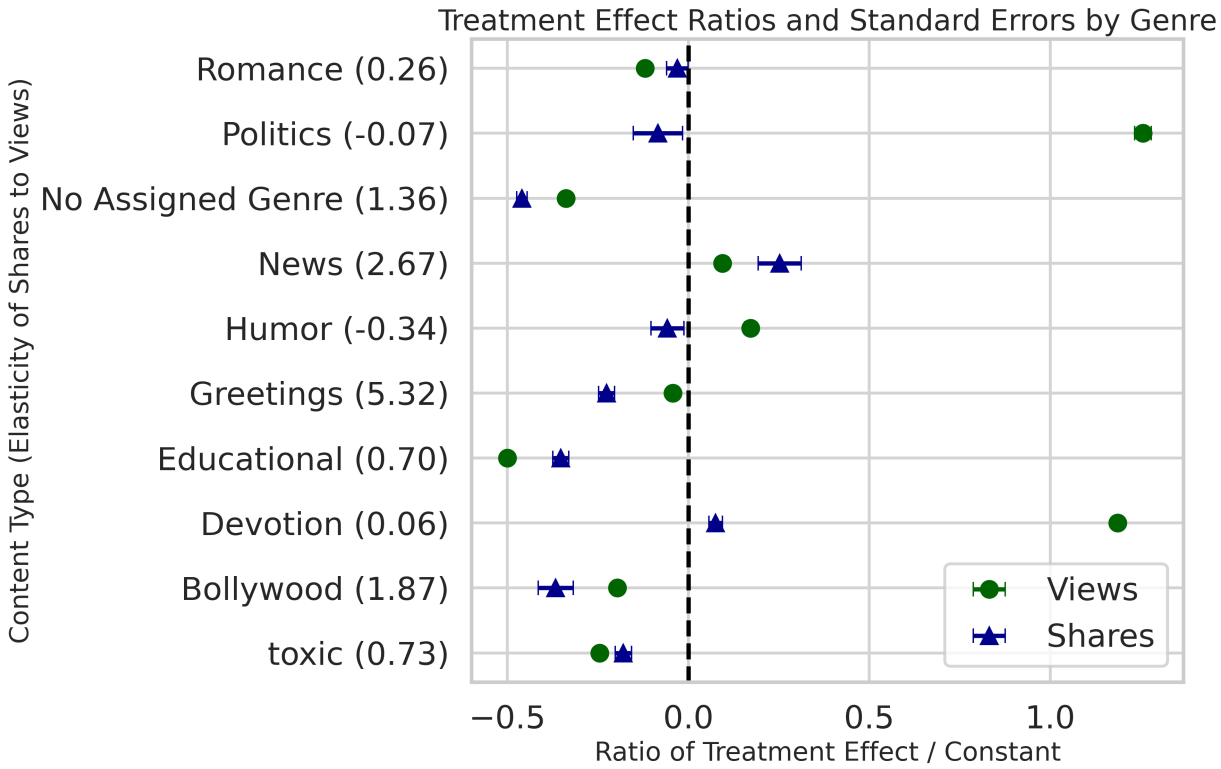
Notes: This Figure shows that the treatment did not affect users' preferences over redistribution, as reflected in the survey data (N = 8,387). This is consistent with the main results that the intervention led to very limited behavioral changes. Users in the random sample survey were asked if they thought 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.18: Substitution with other platforms



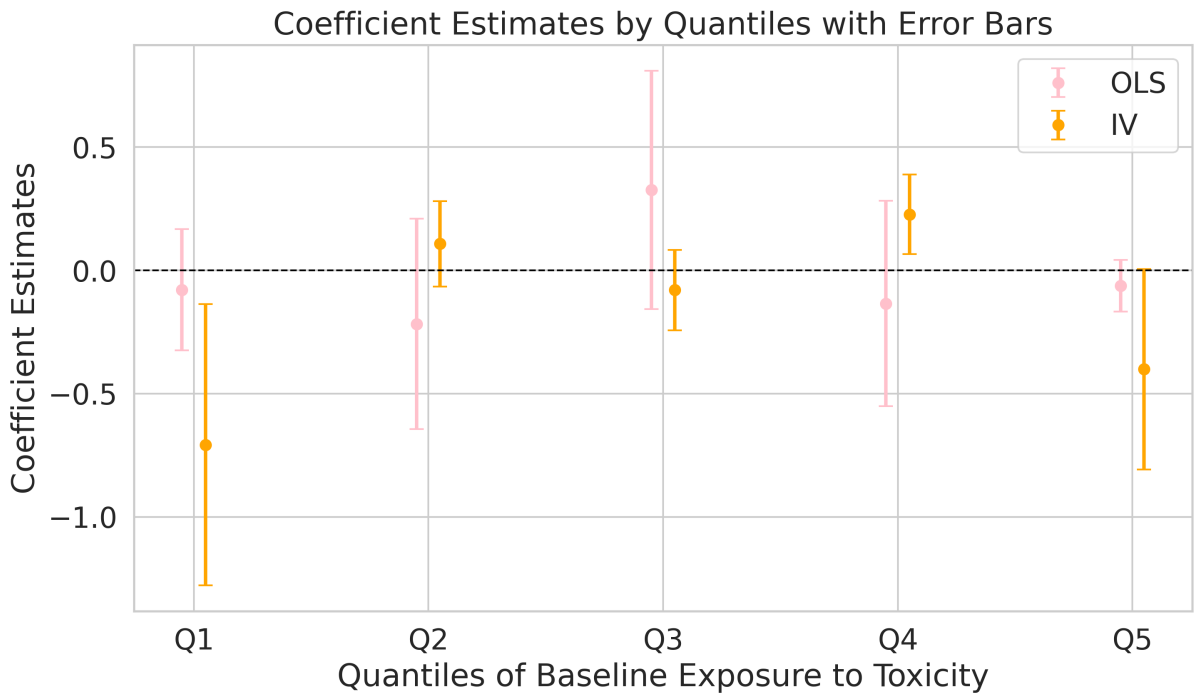
Notes: This Figure shows that users with higher proclivity to toxic content at baseline were more likely to spend more time on other platforms upon being treated (with a p-value of 0.06). A subset of users in the experimental sample were randomly selected for a follow-up survey ($N = 8,387$), and asked 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 D.19: Elasticity of sharing different types of content



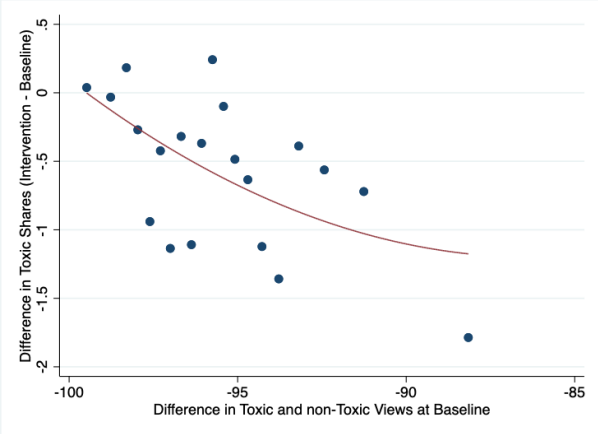
Notes: This Figure shows the elasticity of sharing behavior with respect to exogenous content exposure due to treatment. Each marker shows the ratio of treatment effects to control means with respect to views and shares for different content types. Elasticities are computed by dividing the blue marker (triangle) with the green marker (circle). Relevant elasticities are shown in parentheses next to the labels for each content type. The plot highlights that while sharing behavior for toxic content is inelastic, user behavior for romantic, religious, and educational content is even more inelastic. Standard errors are robust and are computed at the 5% level of significance.

Figure D.20: Testing simplifying assumption in action-signalling model

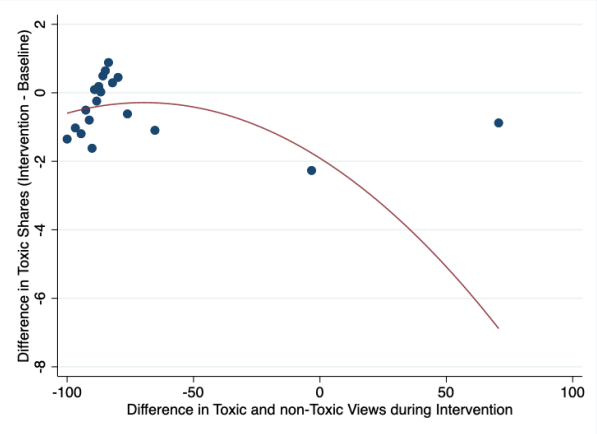


Notes: This Figure shows that I cannot reject the hypothesis that heterogeneous users update their behavior, or are influenced by exposure, at equal rates. This justifies the use of a single parameter θ to capture the rate at which users update their behavior according to the perceived behavior of others, despite their underlying taste for such content. 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. Later, I relax the assumption to estimate the model with malleable toxic users, and mechanical non-toxic users. All regressions were run at the user level, and robust standard errors were computed.

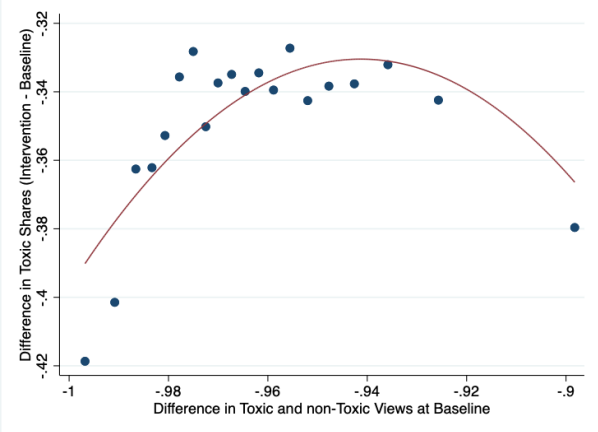
Figure D.21: Structural Estimates and Validation



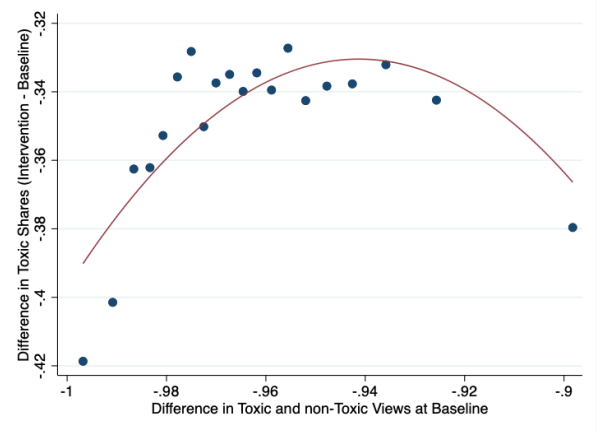
(a) Baseline views and intervention period shares in the treatment group



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



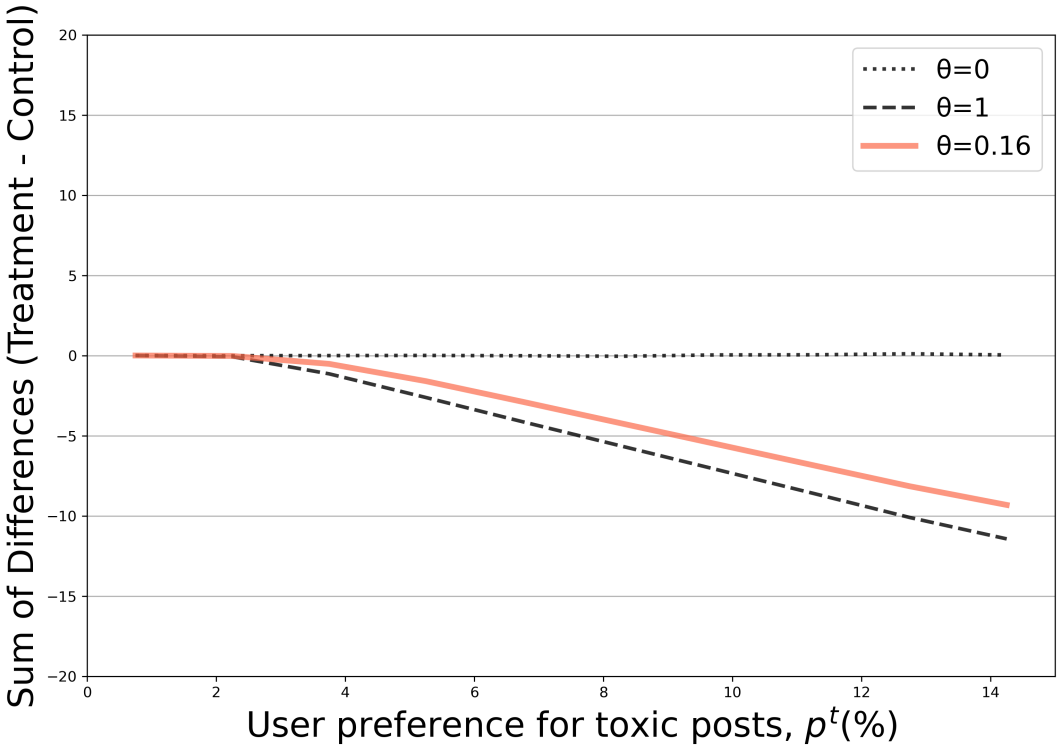
(c) Baseline views and intervention period shares in the control group



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

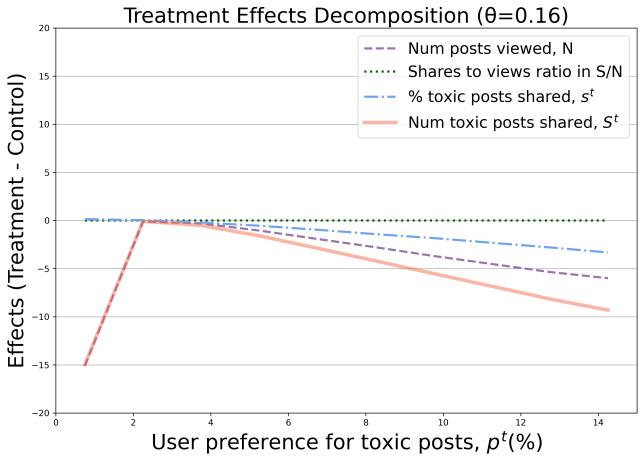
Notes: Panel (a) shows that $\gamma_1 = -\theta$ is negative, and the relationship between differences in toxic shares and toxic views at baseline approximates a linear one, as predicted by the structural model. Panel (b) shows that the relationship between differences in toxic shares (from baseline 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.22: Treatment effects on total number of toxic posts shared for different influence factors, θ

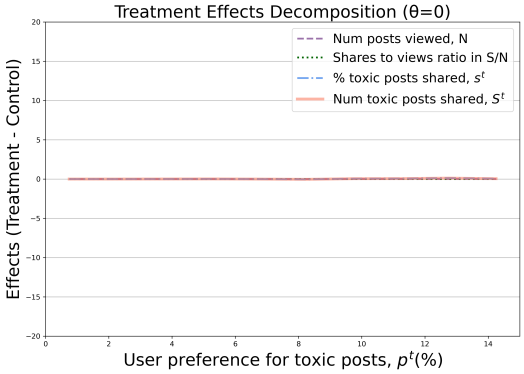


Notes: This figure shows that the simulated treatment effects on number of toxic posts shared is negative for more toxic users, when the rate at which exposure influences behavior is $\theta = 0.16$, as estimated using the structural model and the empirical distributions of various outcomes. This shows that, for the parameter values calibrated using the method of matching moments (See Appendix I.5 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.

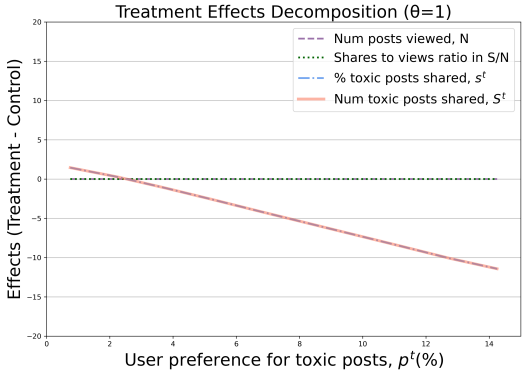
Figure D.23: Decomposition of treatment effects, in different updating regimes



(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)
Mean Dep. Var. in Control	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. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Table E.2: 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 $\mathbb{1}_i(\text{left platform = yes}) = \beta_0 + \beta_1 D_i + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \sum_c \beta_{1c} D_i \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where $\mathbb{1}_i(\text{left platform = yes})$ 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. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Table E.3: User characteristics correlated with the probability of sharing toxic content during the intervention period

Variable	Coefficient	Interaction Coefficient
Treatment Effect	0.054	N/A
	(0.330)	N/A
Num Toxic Posts Viewed	0.058***	-0.020*
	(0.004)	(0.008)
Number of Views (Baseline)	0.000***	-0.000**
	(0.000)	(0.000)
Number of Shares (Baseline)	0.004***	0.002*
	(0.001)	(0.001)
Number of Logins (Baseline)	-0.113***	0.007
	(0.008)	(0.017)
Toxic Views (Baseline)	0.937***	0.024**
	(0.005)	(0.009)
Toxic Shares (Baseline)	0.006*	-0.007
	(0.003)	(0.004)
User Gender	0.056	-0.011
	(0.030)	(0.065)
Days since account created	-0.000***	0.000
	(0.000)	(0.000)
User Age	0.000	-0.002
	(0.003)	(0.006)
Proportion content viewed on weekends	0.152*	-0.167
	(0.065)	(0.136)
Proportion content shared during daytime	0.396***	0.222
	(0.053)	(0.115)
Share of views in Bollywood Genre	-0.835***	0.270
	(0.164)	(0.352)
Share of views in Devotion Genre	-0.092	0.107
	(0.124)	(0.267)
No Assigned Genre	-1.813***	-0.973
	(0.333)	(0.592)
Share of views in Greetings Genre	1.259***	0.195
	(0.100)	(0.220)
Share of views in Humor Genre	-2.244***	-0.420
	(0.397)	(0.810)
Share of views in News Genre	-3.960***	0.017
	(0.239)	(0.555)
Share of views in Politics Genre	-2.491	-0.741
	(1.284)	(3.140)
Share of views in Romance Genre	-0.707***	0.254
	(0.155)	(0.329)

Notes: This Table shows the observable characteristics correlated with the probability of being exposed to more toxic content at baseline, irrespective of treatment assignment. This also reiterates that treatment assignment was balanced in baseline exposure to toxic content. Standard errors are robust at user level are in parantheses. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Table E.4: Testing simplifying assumptions on sharing behavior

	(1) Total Toxic Posts Shared	(2) Total non-Toxic Posts Shared
Total Toxic Posts Viewed	0.012***	
	(0.001)	
Total non-Toxic Posts Viewed		0.011***
		(0.001)
Mean Dep. Var. in control group	1209.2***	34.30***
	(47.54)	(1.465)
N	63041	63041

Notes: This Table provides evidence that the consumption value from sharing both toxic and non-toxic content is equal, which allows the simplifying assumption that each user has the same θ with respect to toxic and non-toxic content. The coefficient estimates are obtained from stacking regressions of (non-)toxic shares on (non-)toxic views. The statistical test of equality of coefficients could not reject the hypothesis that the coefficients from the two regressions are equal. Robust standard errors are in parentheses. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$

Table E.5: 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 parenthesis. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$

Table E.6: Structural estimates using OLS regressions with treated sample

	Difference in shares (Current - Baseline)
Difference in views at Baseline	-0.085*
	(0.039)
Constant	-8.486*
	(3.745)
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. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table E.7: Validating structural estimates using OLS regression with control sample

	(1)	(2)
	Difference in shares (Current - Baseline)	
Difference in views at baseline	0.312**	
	(0.046)	
Difference in views		-0.060**
		(0.004)
Constant	-3.953	-38.561**
	(4.414)	(0.355)
N	168773	168773

Notes: Dependent variable is differences in differences between probability of sharing toxic and non-toxic content, between intervention period and baseline, for control users. The explanatory variables are constructed by averaging differences between proportion of toxic and non-toxic posts viewed by control users. θ estimated in the control sample, in Column (1), is biased upwards. 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		Prop of 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		Prop of 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.2.

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
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
Prop of Toxic Posts Views	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
Prop of Toxic Posts Views	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 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.

G.1 Timing

First, nature randomly assigns user specific parameters $\{\alpha, \beta, p^t, \theta\}$. These parameters not only reflect how a user values consumption of different types of content, but also how she values conformity with peers. That is, nature assigns user tastes for viewing and sharing, β and α respectively. The utility weight user places on conforming with social norms, i.e. θ, is also realized. Later, I show that θ is effectively the rate at which users update their behavior or are influenced by the content they view.

Second, the platform optimizes its ad-revenue by algorithmically assigning content that users are more likely to stay and engage with. Assuming there are two types of posts on the platform, toxic and non-toxic, the algorithms' choice variable in this model is q^t only, with $q^n = (1 - q^t)$. The algorithm chooses these probabilities to maximize the total number of posts viewed by each user, N. Next, for the given assignment probabilities, the user decides the total number of posts she will view N, or the total time she will spend on the platform upon observing the assignment probabilities. The user is thought to learn the distribution of content recommendations that the algorithm would make, so that the choice of N determines the expected number of posts of each type that she views.

The user responds to the realization of the number of toxic and non-toxic posts viewed, through her engagement behavior, i.e. by sharing a viewed post. That is, the user chooses the total number of posts to share, S, which offers the user with some consumption utility, and also scales the behavioral response. Finally, the user chooses the fraction of shared posts that are toxic, $s^t = S^t/S$, in order to maximize utility, for given exposure and sharing decision.

G.2 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 G.1. *For a utility maximizing agent i,*

$$s_{i,\tau}^t = \left(q_{i,\tau}^t\right)^\theta \left(p_i^t\right)^{1-\theta} \tag{6}$$

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 $s_{i,\tau}^r$) from the utility maximization problem stated above in (**??**). □

The optimal sharing strategy is a combination of user's own tastes and the content she is shown, $q_{i,\tau}^t$, 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_i^t. The user also decides the number of non-toxic posts she will share, if any, upon viewing posts in their feed.

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

$$S_{i,\tau} = \frac{1}{2(\eta+\alpha)}\left[2N_{i,\tau}\alpha - \delta\theta(1-\theta)\left(\left(\log p_i^t\right)^2 - 2\log q_{i,\tau}^t \log p_i^t + \left(\log q_{i,\tau}^t\right)^2\right)\right] \tag{7}$$

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 $s_{i,\tau}^t$ 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 G.3. *For a utility maximizing agent i,*

$$N_{i,\tau} = \frac{1}{2\alpha\eta}\left[\beta(\alpha+\eta) - \delta\alpha\theta(1-\theta)\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2\right] \tag{8}$$

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 G.1 and G.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 **??** 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 G.4. *For a utility maximizing agent i,*

$$S_{i,\tau} = \frac{1}{2\eta}\left[\beta - \delta\theta(1-\theta)\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2\right] \tag{9}$$

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 (8) into the optimal sharing function in (9). 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 (8) to the algorithm, which in turn optimally chooses $q_{i,\tau}^t$ to maximize advertisement revenues.

Lemma G.5.

$$q_{i,\tau}^t = p_i^t \tag{10}$$

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 (8), by choosing $q_{i,\tau}^t$ optimally. The same result follows if the algorithm's objective is defined more broadly, choosing $q_{i,\tau}^t$ to maximize $N_{i,\tau}^t$, or $S_{i,\tau}^t$, or some linear combination of the two. This is because the number of posts viewed and shared is decreasing in $\left(\log\frac{q_{i,\tau}^t}{p_i^t}\right)^2 \geq 0$, which equals zero when $q_{i,\tau}^t = p_i^t$. □

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.[41] 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.

H Proofs for Theoretical Framework

H.1 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 G.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. □

[41] 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).

H.2 Proof of Proposition 3

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

$$\frac{\partial^2 s^t_{i,\tau}}{\partial p^t_i \partial \bar{q}^t} \geq 0$$

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 G.1 shows that,

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

Then, we can see that

$$\frac{\partial^2 s^t_{i,\tau}}{\partial p^t_i \partial \bar{q}^t} = \theta(1-\theta)\left(q^t_{i,\tau}\right)^{\theta-1}\left(p^t_i\right)^{-\theta} \geq 0$$

for $\theta \in [0,1]$, and $q^t_{i,\tau}, p^t_i \in (0,1)$. □

H.3 Proof of Proposition 4

User i with $N_{i,\tau}, S_{i,\tau} > 0$, is said to behave 'mechanically' when $\theta = \beta = \eta = 0$. That is, when $\theta = \beta = \eta = 0$, the elasticity of the proportion of toxic posts shared with the respect to the proportion of toxic posts viewed is 1.

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

$$\max_{s^t, S, N} = \alpha(N-S)^2 - \delta S \left(\log \frac{q^t}{p^t} \right)^2 - \eta S^2 \tag{11}$$

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, i.e. $\beta = 0$, we have, $N^t_{i,\tau} = q^t_{i,\tau} N$. Therefore,

$$\frac{S^t_{i,\tau}}{S_{i,\tau}} = \frac{N^t_{i,\tau}}{N_{i,\tau}} = p^t_i = q^t_{i,\tau} \tag{12}$$

Then the treatment implies that,

$$\frac{\partial s^t_{i,\tau}}{\partial \bar{q}^t} = \frac{\partial v^t_{i,\tau}}{\partial \bar{q}^t} = 1 \tag{13}$$

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

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

□

H.4 Proof of Lemma B.1

Lemma H.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 Proposition 4 which shows that $\theta > 0$. Therefore, θ is not estimable from this relationship, in the sample of control users. □

H.5 Proof of Proposition 5

For some updating parameter θ, and treated user i, the change in ratio of toxic-shares to non-toxic shares from the baseline is a function of the log-odds ratio of the proportion toxic posts viewed at baseline. That is,

$$\log\left(\frac{s_{i,1}^t}{s_{i,1}^n}\right) - \log\left(\frac{s_{i,0}^t}{s_{i,0}^n}\right) = (1+\theta)\log\left(\frac{\bar{q}^t}{\bar{q^n}}\right) - \theta\log\left(\frac{v_{i,0}^t}{v_{i,0}^n}\right)$$

Proof. The optimal sharing function for treated users is given by

$$s_{i,1}^t = \left(v_{i,1}^t(\bar{q}^t)\right)^\theta \left(w_i^t\right)^\mu \left(s_{i,0}^t\right)^{1-\theta} \tag{14}$$

$$s_{i,1}^n = \left(v_{i,1}^n(\bar{q}^n)\right)^\theta \left(w_i^n\right)^\mu \left(s_{i,0}^n\right)^{1-\theta} \tag{15}$$

The steady state condition is

$$s_{i,0}^t\left(v_{i,0}^t, w_i^t\right) = s_{i,1}^t\left(v_{i,1}^t(q_{i,1}^t), s_{i,0}^t, w_i^t\right) \tag{16}$$

$$s_{i,0}^n\left(v_{i,0}^n, w_i^n\right) = s_{i,1}^n\left(v_{i,1}^n(q_{i,1}^n), s_{i,0}^n, w_i^n\right) \tag{17}$$

This simplifies to

$$s_{i,0}^t = \left(v_{i,0}^t\right)\left(w_i^t\right)^{\frac{\mu}{\theta}} \tag{18}$$

$$s_{i,0}^n = \left(v_{i,0}^n\right)\left(w_i^n\right)^{\frac{\mu}{\theta}} \tag{19}$$

By design of the experiment, the treated users view a constant proportion of toxic content during the intervention period. Reiterating,

$$v_{i,1}^t = \bar{q}^t \quad \text{and} \quad v_{i,1}^n = \bar{q^n} \tag{20}$$

Then, define $\log\left(\frac{s^t_{i,1}}{s^n_{i,1}}\right)$ as the log-odds of sharing toxic content. Plugging values from (14), (18) and (20) into the definition of the log-odds ratio of sharing toxic content in (**??**), so that

$$\log\left(\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)}\right) - \log\left(\frac{s^t_{i,0}(1)}{s^n_{i,0}(1)}\right) = \theta \underbrace{\log\left(\frac{\bar{q}^t}{\bar{q^n}}\right)}_{\text{constant}} - \theta \underbrace{\log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right)}_{\text{by eqm. cond. (18)}}$$

which is the required expression in the stated proposition. □

I Details of Structural Estimation

This Appendix provides supplementary information on the structural model that enables estimation of some key parameters of interest. First, I discuss why the setting requires a model to estimate θ, and the reasons that standard estimation strategies are not applicable.

I.1 The Necessity of Structure

Since exposure and engagement with toxic content are endogenous variables due to the personalization algorithm, the demand and supply factors that drive problematic behavior cannot be disentangled using the control group data. This is because the algorithm is trained on some underlying user preferences that also determine user behavior, but are not observed by the researcher. This is the main identification problem that the experiment solves. The experiment replaces algorithmically generated content recommendations with a random draw of posts to identify user 'demand' for different types of content.

However, the reduced form relationship between toxic exposure and engagement across treatment and control does not identify the main mechanism of interest: the influence of *exposure* to toxic content on engagement behavior with respect to such content. Without additional assumptions, the experimental variation cannot distinguish between the components of user behavior: innate tastes for toxic content and behavioral responses due to preference for conformity. This is because a draw of posts is picked randomly for the treated users each day, and the Law of Large Numbers implies that the average proportion of toxic posts viewed by treated users is constant over time, at the average probability of being assigned toxic content in the control group.

The random algorithm assigns content in a way that is independent of both these components of user behavior, because the assignment probabilities are drawn from the control distribution of assignment probabilities, *each day*. The Central Limit Theorem (CLT) predicts that due to these daily random draws, the probability of assigning toxic posts for the treated group converges to a normal distribution centered at the average probability of assigning toxic posts in the control group. Further, the variance of the assignment probabilities in treatment is much lower than the control (by a factor of $\sqrt{|\{i|D_i = 1\}|}$) (Hayashi, 2011). In fact, exposure to a particular type of content is almost always constant among

treated users, rendering the relationship between exposure and engagement unidentifiable in a regression of engagement on exposure, in the treatment group.

The concentration of the treatment embeddings around the mean of the control embeddings is also demonstrated in a simple personalization algorithm that is trained on simulated engagement data (See Section 3 and Appendix **??**). I further belabor this point about the absence of a straight forward identification strategy using the reduced form relationship between toxic viewing and sharing in the treatment group. Consider a difference-in-difference estimator to identify the behavioral chain of effects from the change in toxic views because of the intervention, to the change in toxic shares

$$\mathrm{E}\left[\log \frac{s_{i,1}^t}{s_{i,0}^t}\middle| D_i = 1\right] - \underbrace{\mathrm{E}\left[\log \frac{s_{i,1}^t}{s_{i,0}^t}\middle| D_i = 0\right]}_{=0 \text{ in steady state}} = \theta\left(\underbrace{\mathrm{E}\left[\log \frac{v_{i,1}^t}{v_{i,0}^t}\middle| D_i = 1\right]}_{\lim_{i\to\infty} = \text{ constant}} - \underbrace{\mathrm{E}\left[\log \frac{v_{i,1}^t}{v_{i,0}^t}\middle| D_i = 0\right]}_{=0 \text{ in steady state}}\right)$$

where, D_i is an indicator for treatment status. This does not identify the main parameter of interest θ, because the algorithm and control users remain in steady state during the intervention period. As a result, differences in toxic views and shares from baseline for the control users is always zero. Moreover, the distribution of toxic views for treated users is constant during the intervention period, as the treatment assignment probabilities approximate the mean in the control distribution, for each treated user. Therefore, θ is not estimable using a standard difference-in-difference approach.

The difference-in-difference estimator is also amenable to incorrect interpretation, if not correctly grounded in economic theory. The DiD specification implies that θ is the change in sharing behavior due to a change in exposure to toxic content. Suppose I incorrectly estimated the DiD estimator using the data, and found a negative coefficient (Table I.1 shows this to be the case). This would imply that exposure influences behavior negatively, so that users who saw more toxic content during the intervention period shared less toxic content. However, this is not the case, as the structural model shows that the influence parameter is positive, and between zero and one.

Table I.1: Faulty estimates from a difference-in-differences model

	(1)	(2)
	Avg toxicity in sharing	Total toxicity in sharing
DiD estimate	-0.041	0.003^{***}
	(0.028)	(0.001)
N	231814	231814

Notes: This table shows that the difference-in-differences estimate of the treatment effect on sharing toxic content is negative, although statistically insignificant. Column (1) shows the effect on the change in the average toxicity of shares, while Column (2) shows the effect on the change in the total number of toxic posts shared. Column (1) is the closest analog to the structural equation estimated, and the results in Column (2) may be biased because more toxic users may be as treated users may also change the number of non-toxic posts they share. Robust standard errors in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

A model-free interpretation of this estimator implies that the treatment promoted a

backlash against the posts treated users were randomly exposed to. The model provides an estimator for the effect of content exposure that is generalizable to all users, by using exposure to toxic content among treated users at *baseline only*. This enables an interpretation of the mechanisms driving the treatment effects, as the model of behavior below is well identified and accounts for unobserved heterogeneity across users. I distinguish between two channels that drive the treatment effect of switching off the content recommendation algorithms: **(1)** users' innate tastes for problematic content (measured by baseline shares), weighted by $1-\theta$, **(2)** influence of users' perception of society's preferences (measured by toxic exposure), weighted by θ.

I.2 Estimation

I estimate the updating parameter θ from the relationship between baseline exposure to toxic content ($v^t_{i,0}$), and the odds of sharing toxic content during the intervention period ($s^t_{i,1}/s^n_{i,1}$) among *treated* users only. This measures the degree of malleability of user behavior, in the face of algorithm's content recommendations. θ is estimated in a linear regression due to the absence of correlation between toxic views in the pre-intervention period and those during the intervention period among treated users, as shown above. However, notice that regression with logs is not independent of the units of measurement, and the presence of zeroes in the data can generate misleading results (Thakral and Tô, 2023). Further, adding a small constant to the logs that may generate estimates that can be incorrectly interpreted (Chen and Roth, 2023).

I approximate the log-ratios in Proposition 5 with first differences to estimate θ using a Taylor Series approximation (Abbott et al., 2001). Define, $A_{i,\tau} = v^t_{i,\tau} - v^n_{i,\tau}$ and $B_{i,\tau} = s^t_{i,\tau} - s^n_{i,\tau}$. Further, $\Delta B_{i,\tau} = B_{i,\tau} - B_{i,\tau-1}$. Then, the following procedure states the estimation strategy implemented in the data.

Proposition 6. *The behavioral effect of exposure, θ is identified in a linear model of treated users responding to exposure to toxic content if*

(SA1) treated users update their beliefs and sharing behavior in accordance with exposure to toxic content as

$$\frac{d\log(s^t_{i,1}/s^n_{i,1})}{d\log(\bar{q^t}/\bar{q^n})} = \theta$$

(SA2) the influence effect, θ, is constant across time and users.

(SA3) users engagement in equilibrium is stable over time $s^t_{i,0} = s^t_{i,1}$ and $s^n_{i,0} = s^n_{i,1}$

(SA4) assignment probabilities are orthogonal to user preferences conditional on user's observed behavior, $\mathrm{E}[q^t_{i,1}(D_i)|w^t_i, s^t_{i,0}, D_i] = \mathrm{E}[q^t_{i,1}(D_i)|s^t_{i,0}, D_i]$

(SA5) users view all the content assigned to them, $q^t_{i,\tau} = v^t_{i,\tau}$ and $q^n_{i,\tau} = v^n_{i,\tau}$, with $q^t_{i,\tau} + q^n_{i,\tau} = v^t_{i,\tau} + v^n_{i,\tau} = 1$

Then, θ can be estimated using the following regression equation

$$\mathrm{E}[\Delta B_{i,1}|D_i = 1] = \gamma_0 + \gamma_1 A_{i,0} \tag{21}$$

where, $\gamma_1 = -\theta$.

Proof. Consider Proposition 5 which gives

$$\log\left(\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)}\right) - \log\left(\frac{s^t_{i,0}(1)}{s^n_{i,0}(1)}\right) = \theta \log\left(\frac{\bar{q^t}}{\bar{q^n}}\right) - \theta \log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right)$$

Then, consider the second order Taylor series expansion of $\log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right)$, around some fixed point $v^t_{i,0} = v^n_{i,0} = \kappa_1$, where κ_1 is some constant. Then,

$$\log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right) = \frac{1}{\kappa_1} A_{i,0}(1) + O(A_{i,0}(1))^2 \tag{22}$$

This gives the right-hand side of the required expression. Similarly, consider Taylor series expansion of the following term around $\frac{s^t_{i,1}}{s^n_{i,1}} = \frac{s^t_{i,0}}{s^n_{i,0}}$

$$\begin{aligned}
\log\left(\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)}\right) - \log\left(\frac{s^t_{i,0}(1)}{s^n_{i,0}(1)}\right) &= \log\left(\frac{\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)}}{\frac{s^t_{i,0}(1)}{s^n_{i,0}(1)}}\right) \\
&= \log\left(\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)} - \frac{s^t_{i,0}(1)}{s^n_{i,0}(1)}\right) \\
&= \frac{1}{\kappa_1} \Delta B_{i,1}(1)
\end{aligned}$$

which gives the required expression on the left-hand side, as the constant κ_1 cancels out from both sides of the relationship of interest, due to the constant term. □

So, θ measures the behavioral effect of current exposure to toxic content as it gives the rate at which users update their behavior in line with lower exposure to toxic content. Intuitively, θ measures user preference for conformity with societal tastes, that are reflected on user feeds through q^t. Subsequently, θ is the influence parameter, measuring the effect of social norms relayed to a user through exposure to content feeds. This is true in the sample of treated users where, $-\gamma_1 = \theta$ is estimable. Then, $1 - \theta$ provides the appropriate elasticity of user behavior during intervention with respect to user behavior at *baseline*, or with respect to users' inherent tastes for such content.

This model provides the machinery to identify estimates of the stickiness or malleability in human behavior, in addition to the correct interpretation of these results. Furthermore, this interpretation is generalizable for users in both treatment and control groups. Therefore, the model arrives at a non-standard estimation strategy for an important parameter, that cannot be estimated on a social media platform that is typically in steady state. I exploit unique features of this setting and the structural model to estimate parameters that may not be estimated or correctly interpreted using a standard difference-in-difference approach.

I.3 Measurement Error

The estimation strategy above shows that θ, or the influence of exposure to toxic content, is measured using a sample of users in the treatment group, whose exposure to content was completely random. The difference in odds of sharing between the two time periods corrects the bias induced by the omitted variable: users' unobserved preference for *sharing* content. Still, features of the platform and 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 $q^t_{i,0}$. Similarly, each non-toxic post viewed is assumed to be a Bernoulli trial with probability $q^n_{i,0}$. 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 and non-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, $v^{t*}_{i,0}(1)$ and $v^{n*}_{i,0}(1)$ denote the true proportions of toxic and non-toxic content viewed respectively, that are observed with measurement error in the data.

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

where, $ev^t_{i,0}$ and $ev^n_{i,0}$ denote the measurement error in the proportion of toxic and non-toxic content viewed respectively. In general, assume that $Cov(v^{t*}_{i,0}(1), ev^t_{i,0}(1)) = 0$ and $Cov(v^{n*}_{i,0}(1), ev^n_{i,0}(1)) = 0$. 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,

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

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 $half1$. Similarly, $v^{\frac{t}{2(+)}}_{i,0}$ 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, per user, are uncorrelated to each other, 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,

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

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

Proposition 7. *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 ($half2$) by a user in a session.*

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

$$v_{i,0}^{\overline{2(-)}^{t}}(1) = v_{i,0}^{\overline{2(-)}^{t*}}(1) + ev_{i,0}^{\overline{2(-)}^{t}}(1)$$
$$v_{i,0}^{\overline{2(+)}^{t}}(1) = v_{i,0}^{\overline{2(+)}^{t*}}(1) + ev_{i,0}^{\overline{2(+)}^{t}}(1)$$

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

Note the first stage regression using $half2$ as the instrumental variable,

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

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

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

Therefore, the IV approach eliminates measurement error, due to the exclusion restriction stated in (AME). The same strategy is applied to all users with regards to non-toxic content as well. This shows that the IV estimation strategy only depends on the true distribution of the main explanatory variable. □

The OLS estimates suggest that this strategy does not account for measurement error, indeed the OLS strategy produces estimates that are biased towards zero. The IV estimates of θ show that a significant portion of user behavior is determined by user tastes. This implies that user behavior is not malleable with respect to exposure to new information.

I.4 Validation of structural estimates

The structural estimates show that users largely follow their old behavioral patterns, and that behavior is barely malleable according to new exposure to toxic content. 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.[42] This model correctly estimates the updating-behavior only for treated users, because for

[42] This is true for toxic-type treated users.

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.

In the case of control users, exposure in the baseline period is related to engagement with toxic content during the intervention period through two channels, **(1) Direct:** User behavior is correlated across time, and **(2) Algorithmic:** Feed-ranking algorithms that expose users to toxic content to maximize engagement using prior user behavior. Since both these channels are correlated in the steady state, by design of the algorithm, the said relationship cannot be estimated in the sample of control users.

θ estimated using the control sample would be biased upwards as the omitted variable $(v^t_{i,1}/v^n_{i,1})$ is correlated with both the sharing behavior $(s^t_{i,1}/s^n_{i,1})$ as well as the exposure at baseline $(v^t_{i,0}/v^n_{i,0})$ in the main estimating equation (21). This is true if there were sufficient variation in equilibrium sharing behavior across the two time periods, as control users are always in steady state. This relationship is not estimable when the outcome is the difference in sharing behavior between the two time periods $(s^t_{i,1}/s^n_{i,1} - s^t_{i,0}/s^n_{i,0})$, in order to account for unobserved heterogeneity in tastes among users. The steady state condition implies that the ratio of shares at baseline and during the intervention period is equal.

Therefore, estimates that employ the control sample are expected to be distinct from the main estimates above. Table E.7 and Figure D.21c show that this is indeed the case. This validates the main estimation strategy because the estimates from the same exercises using the distinct samples of treated and control users yield very distinct results. Additionally, Figure D.21b 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.

I.5 Calibration

After uncovering the measurement error corrected estimates of the influence parameter θ, the model is calibrated using the data to estimate the parameters. This is helpful in understanding the extent to which the model is able to capture the underlying mechanisms of user behavior, and also to analyze the counterfactual distributions of treatment effects, under different possible values of θ. The model decomposes the contribution of each of these channels in driving the treatment effect, and offers insight into the effectiveness of the intervention, had users been more or less malleable to the content they were exposed to (i.e. for different values of θ).

The sample of treated users generates an estimate of $\theta = 0.16$, after correcting for the measurement error. I match moments of the empirical distribution of various outcomes, with the distributions simulated by the model, where θ is set to 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, as shown in section G.

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

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

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

where, v^t, the proportion of viewed posts that are toxic, is the empirical analog of q^t, the assignment probabilities. Then, the moments of these functions are computed over some distribution of v^t, given by some density function $f(v^t)$. The moment conditions for users with lower proclivity to toxic content is given as,

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

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

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

where, m^t 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,

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

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

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

I use numerical integration methods to evaluate these integrals, assuming $v^t \sim EVT1$. Subsequently, the empirical moments are matched with the simulated moments. The objective is to minimize the distance between the empirical and simulated moments, using the six moment conditions, given by equations (26) to (29). 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).

I.6 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 (Akerlof and Kranton, 2000; 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. This is supported by the survey evidence, as demonstrated before.

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. This is because a standard difference in difference approach cannot identify the influence of exposure on sharing toxic content. The influence of exposure on sharing behavior is not estimable in the control group, because both views and shares are in steady state. Further, I cannot estimate the influence of exposure on sharing behavior in the treatment group, because the intervention does not provide sufficient variation in the exposure to toxic content. This is because treated users view the average user's feed, which is constant during the intervention period. I provide a detailed discussion of the implications of the Central Limit Theorem on variation in exposure among the treated in Appendix I.

I.6.1 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)** action-signalling utility from sharing is equal for both types of content, i.e. $\theta^t = \theta^n = \theta$, **(3)** user behavior in the action-signalling model is updated at some constant rate θ across all users, **(4)** 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 (Train, 2009). 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 simplifying assumption, as I observe whether the signalling value from sharing toxic and non-toxic content is equal. This assumption is validated in Table E.4, where I test for the equality of coefficients using stacked regressions. The two regressions estimate the relationship between toxic views and toxic shares, as well as non-toxic views and non-toxic shares. I cannot reject the hypothesis that the coefficients from these regressions are equal.

I test the third assumption, that is, the constant effects with respect to the rate of updating user behavior in the action-signalling model. Figure D.20 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).

I.6.2 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 E.5 shows that I cannot reject the hypothesis that $\delta_1 = 1$, in the measurement error corrected case.