Algorithmic Drivers of Online Behavior: Evidence from a Large-Scale RCT in India

Aarushi Kalra*

Work in Progress

July 1, 2024

Abstract

This paper studies the drivers of engagement with harmful social media content, assessing the role of recommendation algorithms in reinforcing existing behavioral patterns. A large-scale experiment replacing personalization algorithms with random content delivery for more than a million users on a prominent Indian platform reduced overall usage, underscoring the algorithms' importance in optimizing engagement and advertising revenue. The intervention moderated the content feed of more extreme users, exposing them to content typically viewed by an average user. Surprisingly, the estimated probability of 'sharing' harmful anti-minority (or toxic) content increased by 8 pp among the treated users, despite a 9 pp decrease in the proportion of toxic posts viewed on average. The inelastic sharing of toxic content implies users actively engage with preferred posts, regardless of algorithmic recommendations. These results are consistent with a behavioral model where users 'share' content to signal their true characteristics and their conformity with the behavior of other users in their network. I estimate this model to show that a 1% reduction in new exposure to toxic content through the algorithms decreases 'sharing' of toxic posts by influencing user behavior only by 0.16%, such that the remaining variation in engagement is explained by user tastes. Therefore, regulatory measures aimed at reducing engagement with problematic content may have limited effectiveness, as user preferences significantly contribute to the spread of harmful content.

*Department of Economics, Brown University, Providence, RI, 02906 (email: aarushi_kalra@brown.edu). This version of the paper is currently preliminary and incomplete. This study was preregistered on the AEA RCT registry, ID AEARCTR-0010933. This project was generously supported by the National Science Foundation Dissertation Research Improvement Grant, Weiss Family Fund for Research in Development Economics, and Orlando Bravo Center for Economic Research.

1 Introduction

The Information and Communication Technology (ICT) sector in India constitutes over 13% of the country's GDP, and this contribution is expected to increase to 20% by 2025 (ITA, 2024). Firms in the ICT sector, including social media, OTT, and online shopping platforms, heavily invest in developing and improving feed-ranking technologies– machine learning algorithms designed to maximize usage by setting the order in which content is presented to users on various platforms. The personalization algorithms increase the time users spend on platforms by reducing search costs of online content (Goldfarb and Tucker, 2019). Such cutting edge, albeit opaque technologies enhance firm profitability by drawing advertisers to the platforms in large numbers (Gentzkow and Shapiro, 2010).

However, as an average person spent 151 minutes on social media platforms each day in 2022, increased social media usage has amplified concerns about the harms caused by these black box algorithms (Statista, 2023; Allcott et al., 2020). Social media usage is speculated to worsen mental health outcomes, and even radicalize users because online political divisions are thought to be deepened by these algorithms (Guess et al., 2023a; Agan et al., 2023; Van Bavel et al., 2021). This is because recommendation systems may reinforce user beliefs by creating political 'filter bubbles' (Sunstein, 2001).[1] Recommendation algorithms are thought to increase political polarization, and therefore, need to be regulated, even when such regulation makes social media platforms less profitable. In developing countries with weak democratic institutions like India, political divisions on the digital commons manifest as direct attacks by members of the majority group on minorities by propagating toxic anti-minority posts (Garimella and Eckles, 2017).[2]

Lawmakers in the United States Senate as well as the Supreme Court are discussing regulatory frameworks for personalization algorithms adopted in the social media industry, which caters to 300 million users in the US alone. However, the policy implications, both for platforms and users, of regulating recommendation algorithms used in the ICT sector are understudied, especially in the context of developing countries like India. Such policies can only be effective in reducing engagement with harmful content if regulators take into account the fact that consumption is some unobserved combination of user tastes and recommendation algorithms, each reinforcing the other (Hosseinmardi et al., 2024). Therefore, careful empirical work is needed to estimate the effects of customization algorithms in a way that disentangles the effects of user preferences, from the persuation effects of exposure, in driving problematic behaviors online (DellaVigna and Gentzkow, 2010).

This paper addresses three key research questions. First, to what extent are socially undesirable behaviors, such as 'sharing' posts that are hurtful towards minority groups, enabled by exposure to toxic content via personalization algorithms? Second, how malleable

[1]These harms are important to study because political divisions on social media platforms can limit participation of diverse view points in building public discourse, and can be detrimental to the quality of democracy (Acemoglu, 2021). Moreover, these platforms have been shown to cause increased offline violence (Müller and Schwarz, 2021). These phenomenon have most commonly been studied on platforms like Facebook, Twitter or WhatsApp (Aridor et al., 2024; Tucker et al., 2018).

[2]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." See `https://perspectiveapi.com/how-it-works/` for a comprehensive guide.

is user behavior when exposed to automatically personalized social media feeds? Third, what is the cost, in terms of user retention and engagement, of regulating machine learning algorithms to social media platforms? A rich literature in economics has characterized the demand for content from traditional media sources, in addition to measuring the effects of biased media exposure on human behavior.[3] However, the findings of this literature cannot be generalized because unlike broadcast media, social media platforms customize information that users are exposed to, by way of algorithms.

In order to answer these questions, I conducted a large-scale experiment in collaboration with a TikTok-like Indian social media platform, with close to 200 million monthly active users (henceforth called SM), for 11 months in 2023. My intervention effectively 'switched off' the personalization algorithm for close to a million randomly picked treated users, and *exposed them to randomly picked content* instead. Therefore, exposure to content in the treatment group was given purely by chance. This enabled, under certain assumptions, identification of behavioral responses to exposure to various kinds of content via the algorithm. In particular, I study engagement with toxic content when the treatment reduced a user's exposure to such content, where post toxicity was defined as per Google's Perspective algorithm.[4]

Treatment was randomly assigned at the user level, and less than 1% of SM's users were treated.[5] Note that, the recommendation algorithm for control users continued to optimize over various metrics of user satisfaction during the intervention period.[6] SM's typical recommendation algorithm, like that of Netflix, does this by serving users the kind of content they have engaged with in the past. This is operationalized through contextual embeddings that are obtained by factorizing engagement matrices constructed at the user-post level (Athey and Imbens, 2019), as is delineated in greater detail later. These embedding vectors represent some abstract measure of user tastes over social media content as learned by the personalization algorithm. For treated users, these learned embedding vectors were replaced with a random draw of user features or 'taste vectors' in the same latent space. This effectively randomized exposure to posts for treated users during the intervention period.

On average, treatment reduced an average user's exposure to toxic content (measured as the proportion of toxic posts viewed out of all posts viewed) by 8.65 percentage points. Surprisingly, the intervention increased engagement with toxic content (measured as the proportion of toxic posts shared out of all posts shared) by 7.77 percentage points. I also find that about 15% of the treated users left the platform in the 11-month experimental period, and only 5% left immediately after the intervention began, but treated leavers were

[3]See DellaVigna and La Ferrara (2015) for a comprehensive review.

[4]I measure toxicity of content using Perspective's Machine Learning models developed by Google subsidiary Jigsaw, to identify abusive comments (Jiménez Durán, 2022). Perspective has been adopted by organizations like The New York Times to automatically regulate and filter abusive comments (See `https://perspectiveapi.com/case-studies/`).

[5]I could not stratify the randomization procedure, say by geographic location, due to constraints on engineering capacity in SM's data science teams. Note that the experimentation and survey data collection coincided with mass layoffs in the tech sector, which severely impacted SM's economic prospects. In this way, the intervention differed from the stratification strategy proposed in the pre-analysis plan.

[6]Satisfaction is measured using number of logins and number of posts viewed, as well as engagement with posts that are liked, shared, or commented upon between February 10, 2023 and March 10, 2023. The baseline period considered in this study is December, 2022.

indistinguishable from the control leavers along observable user characteristics. I performed various robustness checks to show that the results were not driven by selection of treated users who stopped using the platform.

The inelastic nature of user demand for toxic content suggests that users actively seek out posts aligning with their preferences, even when such content is not readily served to them via the recommendation algorithm.[7] This behavior may be attributed to users placing high value on sharing toxic content when they encounter fewer toxic posts during the intervention, possibly due to diminishing returns from sharing such content. To investigate this phenomenon, I devised a novel statistical test to detect the presence of 'behavioral' effects in the estimated treatment effects. The hypothesis that the treatment 'mechanically' led to a reduction in toxic shares is rejected because the proportional treatment effect on toxic sharing is distinct from the proportional treatment effect on toxic viewing.[8] This means that the marginal returns from sharing toxic content are not constant with respect to the proportion of toxic content viewed.

This statistical test suggests that the returns to sharing toxic posts are, in fact, higher for treated users who were exposed to less toxic content. This could be attributed to users' innate tastes or preferences for toxic content, or because exposure to new content somehow changes user behavior. The results are consistent with a model where users engage with posts not just for their consumption value, but also because they learn about society's tastes for different types of content from the algorithm-generated content feed. In other words, content feeds may expand or contract users' Overton Windows, which are the range of ideas considered acceptable in public discourse (Astor, 2019). Consequently, users may share content to signal their characteristics or their conformity with their own group. The experimental variation intervenes upon this signaling channel by exogenously altering users' perception of society's tastes through the random assignment of content. However, to disentangle the effect of user tastes from the behavioral effect of signaling conformity due to exposure to new content, additional assumptions need to be made.

I build a behavioral model to characterize sharing behavior. In the model, users make sharing decisions by weighing the new information (about society's tastes for tocix content) they receive through the algorithm, against their innate tastes for such content. Then, users and algorithm are said to be in steady state when the probability of the algorithm recommending a toxic post, and the probability of the user sharing the post is expected to be constant over time. This equilibrium condition provides the identifying conditions for behavioral changes with respect to users' baseline behavior (used as a proxy for tastes) and to the algorithm's content recommendations.[9] The estimation strategy derived from the model is non trivial because posts are drawn independently at random from the same distribution on each day for all treated users. As a result, there is constant exposure across treated users on average due to the Law of Large Numbers, resulting in insufficient variation in toxic exposure to directly estimate behavioral responses. Therefore, the effect of random exposure to toxic content does not identify the behavioral effect of content exposure. This

[7]The elasticity of toxic sharing with respect to toxic viewing is estimated to be 0.73, and significantly less than 1.

[8]Here, treatment effect is said to be mechanical when the change in toxic shares is always a fixed proportion of toxic views.

[9]The behavioral effects are parameterized by the rate at which users update their view of the world

is why I cannot use a standard difference-in-difference estimator, but regress differences in toxic sharing on *levels* of toxic viewing.

The structural estimates indicate that the odds of sharing toxic content decrease by 0.16% when the proportion of toxic content viewed decreases by 1%, relative to the proportion of non-toxic content viewed. Therefore, exposure to toxic content via recommendation algorithms is shown to play a significant role in changing online behavior. Additionally, the structural model implies that most of the variation in users' sharing behavior is determined by sharing behavior at baseline. This implies that user behavior in response to content exposure is fairly sticky in the short run.

These estimates are derived after making corrections for measurement error in the independent variable, which is a binary variable indicating if a post viewed at baseline was toxic.[10] This variable is typically measured with error because toxic posts are sampled by the recommendation algorithm from a list of toxic posts according to some binomial distribution. Further, users typically view only a fraction of content assigned to them by the algorithm in a ranked list of posts. This paper makes another methodological contribution by proposing a novel method to correct for this error in the main explanatory variable, i.e. proportion of toxic content viewed at baseline, using a novel instrument. I regress the engagement outcome on half the toxic posts viewed by a user, and instrument this explanatory variable with the other half of toxic posts viewed. This corrects the measurement error due to the exclusion restriction baked into the instrument.

My results have a number of important policy implications for regulating digital technologies to minimize harm and political polarization, especially in developing countries. Any intervention that reduces exposure to toxic content, or increases exposure to other types of content, leads to behavioral changes. A regulator would want to make sure that these behavioral changes do not counteract the mechanical effects of the reduced toxic exposure, to minimize discrepancy between the actual and intended effect of the intervention. I show that while the intervention significantly reduces the proportion of toxic content viewed, a significantly large proportion of new behaviors can be predicted using engagement patterns at baseline. Therefore, any intervention that reduces toxic exposure may not be entirely effective in changing user behavior.

The remaining paper is organized as follows. Section 1.1 identifies contributions of this paper in light of a growing literature on the economics of social media as well as regulation of new black-box technologies and algorithms. Sections 2 and 3 provide details of the context and data respectively. Section 4 delineates the design of the experiment, and provides reduced form estimates that motivate an interrogation of mechanisms via a structural model in Section 5. In this Section, I also model the appropriate measurement error correction when toxic posts viewed are randomly sampled from the list of posts assigned. Finally, Section 6 provides the structural estimates, and Section 7 concludes.

[10] averaged at user level.

1.1 Relevant Literature

Content recommendation algorithms are often accused to boost engagement with undesirable types of content, like hate speech, on social media platforms across the world.[11] However, it is unclear if this phenomenon is driven by solely by algorithmic exposure to toxic content, or by users' pre-existing preferences over such content. This is because the algorithms are optimized to enhance user satisfaction using revealed preferences that the algorithm learns from users' past engagement behavior. On the other hand, user preferences for content could be influenced by the content they view, so that users can convincingly signal that they are engaging with posts that are popular in their network. The main objective of this paper is to understand these human-computer interactions in shaping digital spaces. In doing so, the paper contributes to three strands of the literature.

First, it relates very closely to the recent and major contributions of Guess et al. (2023b) and Guess et al. (2023a) in studying the role of machine learning algorithms employed by Meta in driving political polarization during the 2020 US Presidential Electionsature on the effects of new technologies on political outcomes 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; Fujiwara et al., 2023), mental health (Allcott et al., 2020; Braghieri et al., 2022) as well as targeted attacks on vulnerable communities by non-state actors (Bursztyn et al., 2019; Müller and Schwarz, 2021). This literature mainly focuses on the threats of social media and black box AI technologies in the context of authoritarian regimes, where state actors may be involved in surveillance and pushing pro-government narratives (Beraja et al., 2023a,c).

The literature recognizes the importance of online outcomes like mobile phone usage (Björkegren, 2019), digital addiction (Allcott et al., 2022; Aridor, 2022), news consumption (Levy, 2021), spread of misinformation (Acemoglu et al., 2021), radicalization (Hosseinmardi et al., 2020), and employment of surveillance technologies (Beraja et al., 2023b). There is also an increasing interest in social media's role in engagement with hate speech (Jiménez Durán, 2022; Jiménez Durán et al., 2023), and in regulatory policies to check this trend (Beknazar-Yuzbashev et al., 2022). I contribute to this literature by studying relevant behavioral and political outcomes in the highly fragmented political context of 1.26 billion Indian people, using experimental data that includes millions of social media users.

I do not rely on survey outcomes and instead analyze preferences revealed 'in the wild' through high-frequency engagement behavior of users with millions of posts on SM. Participants in my study were assigned treatment randomly. As a result, I gather data on outcome variables for all users, and not just a self-selected sample of users (who have been commonly asked to consent to participate in surveys in related prior work). Additionally, I observe user behavior for a considerably long time span, i.e. fourteen months. Along with obtaining the administrative data, I also conducted a large-scale and unique experiment to measure behavioral responses to widely adopted technologies. The intervention lasted for 11 months.

Second, this paper relates to a rich literature studying the relationship between media bias and political polarization (DellaVigna and Kaplan, 2007; Gentzkow and Shapiro, 2010, 2011; Chiang and Knight, 2011; Durante and Knight, 2012; Boxell et al., 2022). Even though social media differs from traditional media in some key aspects (most notably, entry cost

[11] `https://www.bbc.com/news/world-africa-67275219`

and personalization of content), existing literature sheds light on how exposure to content changes important aspects of human behavior (Jensen and Oster, 2009; Ferrara et al., 2012). While various channels, like change in social norms or adoption of 'excuses', have been suggested in the literature (Bursztyn et al., 2020a,b; Bursztyn and Jensen, 2017), I show that social media engagement is driven by utility derived from signaling one's type while conforming to one's own own group (Becker, 1991; Besley and Coate, 1992; Bénabou and Tirole, 2011; Ali and Bénabou, 2020; Butera et al., 2022). However, experimental evidence on behavioral outcomes due to (social) media bias and algorithmic content customization is scarce, especially on TikTok-like platforms that currently attracts a large proportion of social media users (Aridor et al., 2024). I add to this literature by studying economic and political outcomes when content exposure is personalized using machine learning algorithms on newer platforms. This is contrasted with the effect of broadcast media in changing behavior.

Finally, I contribute to the burgeoning literature on the economics (and regulation) of black box algorithms (Rambachan et al., 2020; Aridor et al., 2022) by precisely estimating the role of these algorithms in driving engagement with new technologies in potentially harmful ways. These human-computer interactions are especially important to study in the presence of behavioral responses that may have unanticipated or unintended consequences (Björkegren et al., 2020; Hosseinmardi et al., 2024). I study these questions in the context of a very large developing country where these technologies have been introduced only recently, and so robust regulatory frameworks are currently absent. Fairness concerns of algorithms notwithstanding (Kleinberg et al., 2018), there is a growing interest in the impact of algorithms in a world where they have a wide variety of applications (Arnold et al., 2021). 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 (Kleinberg et al., 2022; Agan et al., 2023; Ludwig and Mullainathan, 2021).

My large-scale experiment in the (developing) world's second largest market for social media complements this literature by studying the degree of malleability in human behavior, in the face of algorithms. Using a structural model, I demonstrate how behavioral patterns influence algorithmic recommendations, and how this in turn generates behavioral responses. This is done by disentangling user incentives to use social media platforms to gain consumption utility, as well as public recognition utility from signalling. While social media platforms aggressively use algorithms to retain users on the platform, users are active agents who often seek out the content that aligns with their pre-existing biases. 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.[12] Crucially, user responses may dampen the effects of regulatory policies that could be introduces in developing countries.

2 Background

I lay out details of the context of this study to highlight the importance of behavioral responses to viewing social media posts using new digital technologies. These behavioral responses, in turn, determine content exposure, as feed curation technologies attempt to

[12]Examples of such platforms include Netflix, Facebook, Twitter, Google Ads, and TikTok

learn 'preferences' from human behavior.

2.1 Social Media in India

There has been a sharp increase in political polarization and hate speech on social media platforms around the world (Halberstam and Knight, 2016). The allegations of Facebook whistleblower, Frances Haugen[13], the violent events of January 6, 2021, on the Capitol Hill in Washington DC[14], and Facebook's failure to prevent incitement of violence in Myanmar and Ethiopia[15] have piqued academic interest in the factors that lead to the production and propagation of hateful content on these platforms (Huszár et al., 2022). With over 500 million social media users, India provides a highly relevant context for studying these questions as social media has grave consequences for the socio-political environment in this setting.

India is among the largest markets for online platforms in the world, second only to China.[16] Social media usage has grave consequences for the socio-political environment in this setting, because social media posts are known to have provoked instances of violence, in the form of mob lynchinsg, riots, and hate crimes (Mukherjee, 2020; Arun, 2019). 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 (Mukherjee, 2020; Arun, 2019). This unfortunate incident is by no means an isolated one.

Social media platforms like WhatsApp, Facebook, and Twitter face an unprecedented challenge of moderating content in this large market. This is difficult because these are American companies operating in a vastly different context, where hate speech on social media propagates in a very typical way. 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.[17]

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 carry out their businesses. In agreeing with the government, social media platforms may in fact be biased against opposition parties and pressure groups. This was seen when Twitter decided to suspend various accounts linked with the Farmer's Movement during massive protests against three controversial farm bills that were 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 described which is Modi's Hindu Nationalist Party[18].

Threats to minority communities, stemming from social media usage in India, are speculated to be bolstered by content recommendation algorithms, which are highly sophisticated

[13] https://www.wsj.com/articles/facebook-whistleblower-frances-haugen-says-she-wants-to-fix-the-compan

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

[15] https://bit.ly/3smUoUE

[16] https://www.statista.com/statistics/278341/number-of-social-network-users-in-selected-countries/

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

[18] 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.

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 (Hosseinmardi et al., 2020).

2.2 The Platform

I partner with SM, a rapidly growing social media platform in India, to understand the effects of exposure to extreme content on social media platforms, due to content recommendation algorithms. I study how the nature of online interactions changes with my intervention in SM's rich online social network, which is constructed using sophisticated machine learning algorithms. I assess the contribution of users' existing behavioral patterns with regards to extreme content using data from SM's (close to) 200 million monthly active users.

SM's app features resemble those 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 (Mishra et al., 2022). SM is one of the most popular platforms in the country, where users create and share content in over a dozen regional languages. Content based social networks, such as SM, are centered around topics like Politics, Devotion, and Wishes.

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.1). User interaction in this network is possible only because of the similarity in content that users have shown to engage with. In this sense, SM is very different from X (formerly, Twitter), where the network depends on following and follower patterns. As a result, 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.

2.3 Control (Current) Algorithm

The estimated treatment effect measures changes in engagement with toxic or non-toxic content for treated users. The control group receives standard algorithmic recommendations generated for each user on the platform, while I intervene on the algorithmic recommendations generated for treated users. In this paper, treatment captures exogenous deviations in recommendations from the current algorithm employed by the platform. This is done by randomly drawing content to generate user feeds. Like Netflix, user feeds on SM are usually customized according to preferences revealed via previous engagement with content using Field Aware Factorization Machines (FFM). This setting is therefore, ripe to study the effects of intervening upon some of the most widely used algorithms in the tech industry.

The FFM algorithm essentially tries to learn some approximation of user preferences from their previous engagement with content on the platform. This is done with the objective of optimizing user retention and engagement by serving them content types that 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. This enables the platform to predict which posts should be recommended to users in order to maximize user retention.

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 engagement matrix, or the ranking of various posts for each user. These vector-weights in the space of latent post/ user features are known as embeddings in the semi-supervised machine learning literature (Athey et al., 2021). The user features produced as a result, are latent representations of user behavior revealed in the past. These latent features are represented as a multi-dimensional embedding vector, where each element in the vector represents the estimated weight each user is estimated to put on exactly 32 latent post attributes. The dimensions of these embedding vectors cannot be interpreted in any human-intelligible terms.

3 Data

I use high frequency administrative data on exposure to and engagement with posts by millions of SM users. I also classify the posts that were 'viewed' and 'shared' by my sample of users, to understand the behavioral factors driving engagement with extreme content.

3.1 Administrative Platform Data

I use administrative data from the platform to construct my main outcome variables including, viewership of and engagement with political and toxic content at the user level. Posts on the platform are also characterized by broad tag genres, user generated hash tags, as well as the text associated with the posts using an LDA topic model.

Administrative data provides information on each post that is viewed or engaged with 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 and weekends. This allows me to trace the posts a user was exposed to, and to observe whether they chose to engage with them or not. As a result, I am able to condition engagement statistics on viewership, which is a feature absent from most observational studies that use social media data (Hosseinmardi et al., 2020).

I also observe user characteristics, like their geographic 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 heterogenous treatment effects. These can be important in characterizing heterogeneity in behavioral responses that I observe in the experimental data. The variables and dimensions of heterogeneity used in this analysis were pre-registered with the AEA RCT Registry (Kalra, 2023), and are descriptively summarized in Table 1.

3.2 Toxicity Classification

The administrative data provides user-post level data on viewership and engagement. To measure the main outcome variable, i.e. toxicity of posts engaged with, I further process the text from these posts in the admin data to classify them as toxic. Multilingual hate speech classification is a key part of my research because I study the impact of my intervention on

hatefulness of platform discourse and the content data is in many regional languages (none in English). I use Perspective's machine learning algorithms to identify toxicity in the Hindi text. Toxic content is defined as "a rude, disrespectful, or unreasonable comment that is likely to make one leave a discussion."[19]

Perspective, from Jigsaw/ Google, 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 makes the 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 used in academic research (Jiménez Durán, 2022; Aridor et al., 2024). These models score a phrase "based on the perceived [negative] impact the text may have in a conversation."

I construct a binary variable, called 'toxic,' which takes value 1 when Perspective's toxicity score on a post is higher than 0.2. Figure A.2 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. This validates the need for contextual embeddings for text classification, using Perspective API. In Figure A.3, 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. I find that the separation between the word clusters corresponds to the 0.2 cut-off of 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 in Figure A.4.

4 Experimental Design and Results

I conducted a large-scale experiment in collaboration with SM, to expose a million users to content that was randomly drawn. This 'random algorithm' was in contrast to the usual content feed which is customized according to user behavior in the past. The experiment enables estimation of demand for toxic content, as being distinct from the supply through the algorithms. The objective of this exercise is to identify the role of personalization algorithms, and the underlying mechanisms, in driving engagement with harmful social media posts.

4.1 Treatment

In this experiment, 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. The control group consists of a random sample of users who were exposed to the usual ranked list of posts, where the ranking was determined by user behavior revealed to the algorithm in previous engagements.[20]

[19] `https://perspectiveapi.com/how-it-works/`

[20] I do not use data of the remaining 99% control users, because different users may be subject to different AB tests at a given point in time. The random sample of control users is known to be uncontaminated by a

Content on SM, comprising of image and video-based posts, is created by users and influencers on this platform. 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.[21] The intervention was designed to leave the incentives of content creators unaffected.

4.1.1 General Description of the Intervention

The random draw of posts for the treated user feeds were generated by replacing the algorithmically generated embedding vector (as in Section 2.3) with randomly picked multi-dimensional embeddings. These 'random embedding' vectors generated for treated users were randomly drawn from the distribution of user embeddings that the algorithm generated for control users. That is, the weights which represented the preference weights corresponding to various post attributes for control users, were drawn randomly, for each treated user.[22] The random draws were from the distribution of latent features, encoded in user embeddings, that were generated for control users by the personalization algorithm, on each day.

The multi-dimensional embedding vectors are essentially generated by factorizing engagement matrix at the user-post level, into vectors that characterize both the user and the post in some latent embedding space. For the treatment group on the other hand, embedding vectors are uniformly sampled from an epsilon ball whose centroid is given by the mean embedding in the control group and radius is twice the sum of variances in that vector.[23] In particular, $\boldsymbol{\mu} = \frac{1}{N}\sum_{i=1}^{N} \boldsymbol{x}_i \sigma^2 = \frac{1}{N}\sum_{i=1}^{N} (\boldsymbol{x}_i - \boldsymbol{\mu})^2$, where $\boldsymbol{x}_i$ represents the embedding with bias for user i, and N is the total number of users. For each user embedding, uniformly sample a point from an epsilon ball with the centroid $\boldsymbol{\mu}$ as the center and radius $2 \times$ variance.

$$\boldsymbol{\upsilon} \sim \mathcal{U}(\text{Ball}(\boldsymbol{\mu}, 2\sigma^2))$$

where $\boldsymbol{\upsilon}$ is the new 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 $2\sigma^2$. I show the distributions of a few dimensions of the embedding vectors in the data in Figure A.6, where none of the dimensions are interpreted in human-intelligible terms.

This way of administering the intervention has implications for the distribution of content in the treatment group. I demonstrate the key properties of the content distribution in the treatment group by simulating a simple recommender system with two dimensional embedding vectors. For a randomly generated matrix of engagement for 1,000 control users and 50 posts, I obtain two-dimensional embedding vectors for both the users and the posts using Singular Value Decomposition (Golub and Reinsch, 1971; Wall et al., 2003). Then, I imitate the procedure with which the intervention was implemented, uniformly at random picking embeddings from a bivariate distribution centred at the mean of the embedding vectors generated for control. This process is then repeated each day for a month-long analysis period. In Appendix E, I provide details on how the simulation was implemented.

different intervention being implemented on the platform.

[21] Further, my sample excludes posts that were created more than 60 days before the date that a particular piece of content was served.

[22] Each element in the embedding vector represents latent post attributes that may or may not be intelligible to humans.

[23] As per internal company documents and author's conversations with data science engineers.

First, note that the distribution of treatment embeddings, in Figure 1, approximates a normal distribution centred around the average embedding vector generated by the control group, as would be predicted by the Central Limit Theorem (CLT). Second, the CLT also predicts a smaller spread for the treated embeddings than the control embeddings (Hayashi, 2011). Third, The bivariate distribution of the embedding vectors for treated users is also more concentrated around the mean of the distribution. Finally, and most importantly, the treatment has a greater affect on users with more extreme embedding assignment at baseline. Figure 4 shows that the users with embedding assignments close to the mean at baseline did not see differences in embedding assignment when treated. This is an important characteristic of content distribution, and is later formally stated in terms of treatment intensity as well as intent-to-treat.

4.1.2 Randomization

Treatment was randomly assigned at the user level to 1% of SM's user base, which includes both active and inactive users. Users opt-in to be randomly assigned to the treatment for market research and AB tests conducted that are routinely conducted by the platform.[24] The intervention began on February 10, 2023 and continued till the end of the year (December 31, 2023). Data on relevant outcomes was gathered for the baseline period (December, 2022), intervention period (February to December, 2023) and post-intervention period (January-March, 2024).

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 these users were not subject to any other AB tests/ RCTs. For the analysis in this paper, I use a sample of a quarter million *active* users (62299 in the treatment group, and 189648 in the control) who operate the app in Hindi language. A user is said to have achieved minimum activity if she viewed at least 200 posts during the baseline period. Randomization in the experimental set up provides the independence condition,

$$(Y_{i,\tau}(1), Y_{i,\tau}(0)) \perp D_i \tag{A1}$$

where, D_i is an indicator for treatment, and $(Y_{i,\tau}(0), Y_{i,\tau}(1))$ are potential outcomes for user i, at time τ. I denote the baseline or pre-intervention period as $\tau = 0$, whereas the intervention period is denoted by $\tau = 1$. I verify the validity of randomization in treatment assignment among the sample of active users by assessing balance in observable user characteristics across treatment and control. This is done for the main sample of a quarter million users, that was constructed for analysis. The primary test that I performed is represented by the following regression equation,

$$D_i = \omega_0 + \sum_c \omega_c \mathbf{1}_i(\text{user characteristic} = c) + \varepsilon_i \tag{1}$$

where, treatment assignment (D_i) or user characteristics do not vary over time τ. I consider various observable user characteristics, including gender, state and city of residence, and

[24] My data use agreement does not require additional permissions or consent to be obtained from users, in order to user these data for research purposes.

the week in which a user first created their account. Since the treatment was assigned randomly at the user level, I test that $\omega_c = 0$ for all user characteristics, in order to verify random assignment to treatment. Additionally, I test that all these coefficients are also jointly insignificant, and do not expect to reject this hypothesis.

Table 1 provides estimates for a randomly selected set of attributes, as the full set of user characteristics is too long. This Table provides summary statistics for exposure to and engagement with toxic content at baseline (December, 2022). I am unable to reject the hypothesis that each of these characteristics (individually or jointly) cannot predict treatment assignment. Further, there is balance in behavior at baseline with respect to viewing and sharing all types of posts, across treated and control users. This sets the benchmark for user behavior, and various treatment effects are compared with the control means.

4.1.3 Treatment Intensity

Treatment intensity is defined by the average treatment effect on the total number of toxic posts viewed (sometimes referred as toxic exposure). This outcome measures intensity of treatment for different users because the platform or a social planner can only alter what users are exposed to, by intervening on probabilities with which content is assigned to a user. As before toxicity is defined as a binary variable, that takes value 1 when the toxicity score of a post is at least 0.2.[25] Therefore,

$$ATE^{tView}_{\tau=1} = E[V^t_{i,\tau=1}(D_i)|D_i = 1] - E[V^t_{i,\tau=1}(D_i)|D_i = 0] \tag{2}$$

where, V^t_i denotes total number of toxic posts viewed by user i, as the superscript t henceforth relates to toxic content only. The random exposure to content implies that users who saw more toxic content at baseline would now see a lower number of toxic posts. This is also seen in the recommender system I simulated above, where any one dimension of the embedding vector could encode post or user toxicity.

I have verified that treatment assignment was randomized at the user level. This implies that the sample of treated users is representative of different kinds of platform users in the population. That is, treatment was assigned to toxic or non-toxic users with probabilities equal to the proportion of toxic or non-toxic users in the population, respectively. I define user type (for example, I have called some users toxic and other non-toxic) using their baseline exposure to and engagement with different types of content.

I place both treated and control users into different quantiles of toxicity, depending on the percentage of their feed that was toxic at baseline (December, 2022). Therefore, a user is binned with other users who had similar levels of toxic exposure at baseline (See heterogeneity in users' baseline toxicity in Figure 9). Therefore, each bin represents user type in terms of toxic engagement.

I measure treatment intensity simply as the heterogeneous treatment effect on total number of toxic posts viewed for each group, defined using baseline exposure to toxic content,

$$ATE^{tView}_{g,\tau=1} = E[V^t_{i,\tau=1}|D_i = 1, Group_{i,\tau=0} = g] - E[V^t_{i,\tau=1}|D_i = 0, Group_{i,\tau=0} = g] \tag{3}$$

where, g indexes the group that user i belonged, $g \in \{1, 2, \ldots, G\} \equiv Group_{i,\tau=0}$ arranged in ascending order of the proportion of user feed that was toxic at baseline.

[25] See Section 3 for details on how the cut-off criterion of 0.2 was determined.

The properties of the simulated recommendation algorithm above showed that users who were assigned embedding values closer to the average at baseline were users who faced low treatment intensity upon being treated. Therefore, I expect that users with average levels of toxic exposure to face the lowest treatment intensity, compared to all other groups. Figure 6 shows that the treatment intensity, in terms of exposure to toxic content, is positive for users with low baseline exposure to toxic content. As expected, the treatment effect on number of toxic posts viewed is negative for users with high degree of toxic exposure at baseline (Q5 to Q10). Finally, the average treatment effect on toxic exposure in Q4 group cannot be differentiated from zero.

4.1.4 Intent to Treat

The intent-to-treat with respect to anti-minority and/ or hurtful content is measured as the difference between the total number of toxic posts viewed and the number of toxic posts that were predicted to be viewed based on past usage. That is,

$$ITT_{i,\tau=1}(D_i, V^t_{i,0}) = V^t_{i,\tau=1}(D_i) - E[V^t_{i,\tau=1}|V^t_{i,\tau=0}, D_i = 0, m_i] \tag{4}$$

where, $V^t_{i,\tau=1}(D_i)$ is the number of toxic posts viewed by user i, during the intervention period $\tau = 1$, under treatment assignment D_i. Additionally, the prediction of total number of toxic posts viewed is *conditional* on the total number of posts viewed, m_i, which is assumed to be independent of treatment status, while computing the intent-to-treat. Then, $E_{i,\tau=1}[V^t_{i,\tau=1}|V^t_{i,\tau=0}, D_i = 0, m_i]$ is the predicted number of toxic posts that would be viewed by user i during the intervention period, for given number of toxic posts viewed at the baseline period, $\tau = 0$, if the user were to be assigned to the control group. Therefore, I rewrite $E_{i,\tau=1}[V^t_{i,\tau=1}|V^t_{i,\tau=0}, D_i = 0, m_i]$ as $E_{i,\tau=1}[m_i \cdot v^t_{i,1}(D_i)|V^t_{i,\tau=0}, D_i = 0, m_i]$, where $v^t_{i,1}(D_i)$ is the proportion of toxic posts viewed during the intervention period, $\frac{V^t_{i,1}}{m_i}$.

The intent-to-treat is the difference between the proportion of toxic posts viewed, and the proportion of toxic posts the user *would have viewed* if she was not treated. This ITT measure is built under the assumption that there is no margin of endogenous response in exposure to toxic content, because the total number of toxic posts viewed does not change. As a result, treatment affects toxic exposure only by changing the proportion of toxic posts that are assigned to the treated user.

The intervention was operationalized using random assignment probabilities instead of probabilities customized according to users' past usage. Note that, the treatment intensity differs from intent to treat for treated users, within each heterogeneous group. This is because of endogenous response in total number of posts viewed upon being assigned treatment, that affects the treatment intensity. The intent to treat is measured under the assumption that the total number of posts (of any type) viewed is independent of treatment status.

Figure 6 shows the intent-to-treat for heterogenous users according to baseline exposure to toxic content. I find that the ITT follows the direction of, but is always larger than, the treatment intensity. This is because, on average, treated users reduce the total number of posts they view when the 'random algorithm' changes assignment probabilities for different types of content. The differences between ITT and treatment intensity demonstrate that treated users respond using the margin of total posts that can be viewed. Both these endogenous choices are employed to build a model of user behavior later.

4.2 Outcomes

I study the effect of treatment, which turns off feed personalization via content recommendation algorithms, on measures of satisfaction and engagement on the platform, as well as on socially undesirable behavior. Treatment effect on measures of overall platform engagement, like total number of posts viewed and time spent, helps identify the value of the algorithm for the customers as well the cost of the intervention to the platform. The main objective is to understand the sign and magnitude of the treatment effect on toxic shares, in order to decompose it into its demand-side constituents.

4.2.1 Overall Platform Engagement

Since the content recommendations are tailored according to user tastes, I expect a deviation from personalized recommendation in the intervention to reduce user retention and engagement on the platform. This is because personalization algorithms are thought to decrease user's cost of searching for the content she likes. In reducing ease of content discovery, the treatment is expected to reduce overall engagement with the platform.

I measure the average treatment effect on the total number of times a user logged on to the platform (also referred to as retention), the total time they spent online, and the total number of posts they viewed and shared. The treatment effects in Table 2 show reduced engagement with the platform, and are estimated using

$$\mathrm{E}[Y_{i,1}(D_i)|D_i] = \psi_0 + \psi_1 D_i \tag{5}$$

where, $Y_{i,1}$ denotes the outcomes for user i during the intervention period ($\tau = 1$), and D_i is an indicator for treatment assignment. As a result, the intent-to-treat (ITT) in (5) implies that $\psi_0 = E[Y_{i,1}|D_i = 0]$ and the main coefficient of interest, $\psi_1 = E[Y_{i,1}|D_i = 1] - E[Y_{i,1}|D_i = 0]$, is the average treatment effect, under the unconditional independence assumption (A1).

In Table 2, disengagement in all aggregate measures of platform usage is noted due to negative and statistically significant treatment effects on total number of posts viewed and the time spent on the platform. This means that the intervention was costly for the firm, because an average treated user reduced the total time she spent on the platform by 2.5 hours, compared to an average control user. As a result, treated users saw less advertisements, and this could impact the ad-revenues that the platform attracts. This also has implications for the value for consumers that the algorithm generates.

I find that treated users were more likely to leave the platform (to never login again after leaving) in the first three months of the intervention. However, Table 4 shows that the treated leavers were not different from the stayers along any characteristics observable in the data.[26] That is, I cannot rule out the hypothesis that, in the following regression equation, $\phi_{3,c}$'s are individually or jointly insignificant.

$$\mathbf{1}_i(\text{left platform}) = \phi_0 + \phi_1 D_i + \sum_c \phi_{2,c} \mathbf{1}_i(\text{user characteristic} = c) + \sum_c \phi_{3,c} \mathbf{1}_i(\text{user characteristic} = c) \cdot D_i + \varepsilon_i \tag{6}$$

[26] This is complemented by evidence in Table B.1, which shows balance across treatment and control groups, along observable characteristics, in the sample of stayers.

where, $\mathbf{1}_i$(user characteristic $= c$) represents indicator variables for observable characteristics c of user i, including gender, age, days since account creation, as well as weekday and daytime usage at baseline. I find that the decrease in total views and shares is due to the extensive margin of users who drop-off the platform entirely. That is, I show this in Figure 11, because among the sample of users who continued to stay on the platform after the intervention, there is no change in the total number of posts viewed and shared.

Since treated users experience different treatment intensities depending on their baseline platform usage, I further predict that users with higher treatment intensity were more likely to leave the platform, or disengage from the content on it. Differences in treatment intensities was illustrated through the assigned embeddings in a simulated recommender system, which generated Figure 4. This shows that the users with embedding assignments close to the mean at baseline did not see differences in embedding assignment when treated. As a result, the treatment intensity is highest for users with more extreme embedding assignments at baseline. Such users were, therefore, expected to be more disengaged with the platform at baseline.

For treatment intensity measured in terms of exposure to toxic content, treated users who viewed only one 1 toxic post at baseline were seen to have positive treatment intensity in Figure A.7. On the other hand, users who saw more than one toxic post received intensity of that treatment that was negative. In order to analyze treatment effects on groups of users heterogenous by treatment intensity, I divide the sample of users into 10 quintiles based on their exposure to toxic content at baseline.

Figure 9 suggests that the treatment effect on total number of logins is indistinguishable across heterogeneous users. Further, heterogenous users by baseline toxicity have very similar levels of total posts viewed and shared. However, I did not find any evidence that users who are heterogeneous along the toxicity dimension differ in terms of observable characteristics. This result can be explained by one of the main findings developed in this paper: users engaging with extreme content at baseline seem to seek out the content they like anyway. That is, even when it is harder for a user to find the content she likes, she is more likely to spend time on the platform searching or scrolling past content to get to the posts she wants to engage with.

4.2.2 Toxic Engagement

The objective of this intervention was to understand the role of the customization algorithms in driving socially undesirable behavior, such as online sharing of harmful or toxic content. I analyze this using outcomes such as total number of toxic posts viewed and shared, where toxicity is defined as a binary variable, that takes value 1 when the toxicity score of a post is at least 0.2.[27] Therefore,

$$ATE^{tView} = E[V_i^t(D_i)|D_i = 1] - E[V_i^t(D_i)|D_i = 0] \tag{7}$$

$$ATE^{tShare} = E[S_i^t(D_i)|D_i = 1] - E[S_i^t(D_i)|D_i = 0] \tag{8}$$

where, V_i^t denotes total number of toxic posts viewed by user i, and S_i^t denotes the total number of toxic posts shared by user i. The random exposure to content implies that users

[27] See Section 3 for details on how the cut-off criterion was determined

who saw more toxic content at baseline would now see a lower number of toxic posts. For all such users, I expect to see a reduction in the total number of toxic posts shared because **(1)** these users have a lower number of toxic posts to share from, and **(2)** seeing more diverse content may change their otherwise toxic attitudes.

Figure 7 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, and treated users also shared a lower number of toxic posts in absolute terms. In fact, treatment reduced the proportion of toxic posts viewed by 8 percentage points, but increased the proportion of toxic posts shared by 7 percentage points. This is because even though an average SM user saw less toxic posts, and shared a smaller *number* of toxic posts, this average user shared less of other types of content. The is because users seek out content that they like, even when the algorithm does not serve it to them readily. Therefore, change in overall engagement, or behavioral changes, may explain these results.

Figure 11 shows that among the sample of users who continued to stay on the platform after the intervention, there is no change in the total number of posts viewed and shared. However, the exposure to and engagement with toxic content, both reduce for the average treated stayer. This shows that selection is not driving the main results with respect to socially undesirable behavior. This provides the foundation for the claim that the intervention induced behavioral changes.

As before, the treatment intensity, with respect to number of toxic posts viewed, differs by user behavior at baseline. Therefore, reduction is exposure to toxic content is primarily driven by users who saw more toxic content at baseline. This is demonstrated in Figure 12, where the negative treatment effect on number of toxic posts viewed is largest for users in the highest quantile of the distribution of users viewing toxic content at baseline. Note that for this quantile, the reduction in total number of toxic posts shared is not as large as the reduction in toxic views.

4.2.3 Elasticity

The results demonstrate that users seek out the content they like. In particular, users who engaged with more toxic content at baseline were likely to share a higher proportion of toxic content even when they saw a lower proportion of toxic posts. This implies behavioral change among treated users. I test if the total reduction in toxic shares is 'mechanical' because users share a fixed fraction of toxic posts they view. Mechanical sharing would imply that users always share a fixed fraction of toxic content they view. For this analysis, I define proportionate change in exposure to toxic content,

$$ATE^{pView} = \frac{E[V_i^t(D_i)|D_i = 1] - E[V_i^t(D_i)|D_i = 0]}{E[V_i^t(D_i)|D_i = 0]} \tag{9}$$

where, V_i^t denotes total number of toxic posts viewed by user i. Now, consider engagement behavior (sharing) with respect to toxic content, so that

$$ATE^{pShare} = \frac{E[S_i^t(D_i)|D_i = 1] - E[S_i^t(D_i)|D_i = 0]}{E[S_i^t(D_i)|D_i = 0]} \tag{10}$$

where, S_i^t denotes the total number of toxic posts shared by user i. If users are seeking out toxic content, then the proportionate change in toxic sharing is smaller than that proportionate change in viewing (in absolute terms). That is, this model of user behavior predicts that the elasticity of sharing toxic content is less than 1. Then, the elasticity of engagement with toxic content (in the form of shares) with respect to exposure to toxic content is

$$\epsilon(S_{i,1}^{toxic}, V_{i,1}^{toxic}) = \frac{ATE^{pShare}}{ATE^{pView}} \tag{11}$$

Figure 8 shows that the elasticity of toxic sharing with respect to toxic viewing is less than 1. This figure also shows that user behavior is malleable for some content categories, whereas it is extremely sticky for categories like Devotion. Further, $\epsilon(S_{i,1}^{toxic}, V_{i,1}^{toxic}) = 1$ is the case of mechanical effects under an assumption of constant effects. Intuitively, this means all users always share a fixed proportion of toxic content that they view.

In order to test the hypothesis of mechanical effects, I check if the ratio of toxic shares to toxic views is different across treatment and control.[28] This hypothesis can be directly tested by considering the log transformations of the outcome variables. However, it is unclear if this approach will yield the correct estimates due to a very high proportion of zeroes in the outcome data (Chen and Roth, 2023; Thakral and Tô, 2023). I test this hypothesis of mechanical sharing using a novel empirical test that I develop below.

Mechanical effects imply that the change in sharing behavior with respect to toxic content is a constant proportion of the changes in toxic exposure. Furthermore, for this special case when there are no behavioral effects, the probability of sharing a toxic post s_i is assumed to not respond to treatment assignment. I test for mechanical effects by making three assumptions that are obtained on account of the experimental design, **(M1)** treated users viewed toxic content that was necessarily picked randomly, so that

$$v_i^t(1) \perp s_i \tag{M1}$$

and **(M2)** in the special case that sharing behavior is stipulated to be purely mechanical, suppose that,

$$E[s_i^t | v_i^t(0)] = \pi, \quad \forall i \tag{M2}$$

where, $\pi \in (0, 1)$. That is to say, the conditional probability of sharing toxic content, if user i was assigned to the control group, is just some fixed fraction irrespective of the proportion of toxic posts viewed. Further, this fraction is constant across all users i.

Thus, if users behaved mechanically, the total number of toxic posts shared is a fixed proportion of the total number of toxic posts viewed by all users. Then, the relative change in proportion of toxic content shared should equal the relative change in proportion of toxic content viewed. Finally, **(M3)** makes an assumption of constant effects.

Proposition 1. *Consider a special case of the model so that sharing behavior does not respond to treatment assignment or the exposure to toxic content during the intervention period. Further the probability of sharing toxic content, for given levels of toxic exposure, is*

[28] I use the term mechanical effect as a short hand for the case where the treatment effect on toxic shares is a constant proportion of the treatment effect on toxic views. This is assumed to be the fixed proportion of toxic content that users shared out of the toxic content they viewed at baseline.

constant across all users. Then the proportionate change in sharing toxic content is exactly equal to the proportionate change in toxic content viewed, i.e. $ATE^{pShare} = ATE^{pView}$

Proof. In Appendix. □

The proposition provides a direct way to test the hypothesis that treatment causes shifts in engagement with toxic content only on account of shifts in exposure to toxic content. This implication would be true if users' sharing behavior with respect to toxic content is not responsive to exposure, and is constant across all users. I have already noted that the elasticity of toxic shares with respect to toxic views is less 1. I formally test the non-linear hypothesis of equality of the two ratios in question, and reject it with a χ^2-stat of 9.44 and a p-value of 0.002.

5 Theoretical Framework

Personalization algorithms have been criticized for their role in radicalizing users by creating echo chambers of like-minded individuals who usually engage with toxic content. (**?**). I intervened in these social networks by replacing the personalization algorithm with a random algorithm, where users were recommended a random draw of posts in their social media feeds. This paper is organized around the costs and benefits of this intervention administered to randomly selected users. The costs of the intervention were borne by the platform as it led to a decrease in the time users spend on the platform, and hence a decrease in the revenues generated from advertising. This is because treated users were not served personalized content, and were more likely to see content that they were not be interested in. On the other hand, the intervention offers benefits to society by reducing the amount of toxic content shared. The intervention may indirectly compound these benefits due to behavioral responses of exposing users to different types of content (for example, users who were more likely to engage with toxic posts at baseline were shown more neutral posts, and this may enable such users to change their minds about, say, minority groups). This is because users derive utility from signalling their conformity with social norms, by sharing posts that are closer to their perception of society's tastes for toxic content.

The treatment impacts user engagement with toxic content via three channels. First, in assigning random draws of posts to treated users, the intervention directly reduced the proportion of toxic posts viewed by users who engaged with more toxic content at baseline. This mechanically left such users with fewer options to share toxic content from. Although the intervention marginally increased the proportion of toxic posts viewed by users who engaged with less toxic content at baseline, the average treated users viewed fewer toxic posts. Second, treated users endogenously responded to new content assignment by viewing fewer posts, or spending less time on the platform. This is because the algorithm's recommendations were no longer tailored to their innate preferences. The effect was pronounced for users who were previously engaged with more extreme toxic content (henceforth, toxic users), because they were more likely to be affected by the treatment.

Larger treatment effects for toxic users is implied by the larger differences between users' latent features (computed by the true algorithm) and the average user embeddings that are more likely to be assigned to treated users during the intervention period. The design of the

intervention therefore, bakes in heterogenous treatment intensity across users, and identifies conditional treatment effects for given toxic engagement at baseline. Third, the treatment altered the perception of society's tastes for toxic content, which users infer from the realized sequence of posts that appear in their feed. 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). This change in perception elicits behavioral responses from users, that are in line with the slant of the content they are exogenously exposed to.

The experimental data is sufficient to identify the first channel which altered the supply of content. This results from the fact that the content recommendations are completely independent of user preferences in this experiment. However, I cannot disentangle the second and third channels without imposing additional structure on user behavior. That is, treatment may lead users to disengage with the platform by spending less time on it, but exposure to new types of content may change user behavior by altering their perception of society's tastes. Both these mechanisms are likely to be at play, especially when users receiving utility from signalling their conformity with continuously changing social norms. This action-signalling model (Butera et al., 2022) helps in disentangling users' endogenous response on the total number of posts viewed from the behavioral effect on the proportion of toxic posts shared. This is important because the former provides accurate estimatess of the cost of the intervention to the platform, while the latter predicts the role of behavioral distortions in driving the treatment effect on toxic sharing.

5.1 Micro-Foundations

A user's incentives are modeled in the spirit of Akerlof and Kranton (2000), where the user receives payoffs from projecting a certain image of herself to her peers. The user cares about how her action of sharing a post compares to some linear combination of behavior among the population as well as one's own true characteristics. This provides the micro-foundations to structurally estimate a behavioral model, where the main parameter of interest is the rate at which users update their engagement behavior when exposed to new content that may be slanted (DellaVigna et al., 2012). The distinction between behavioral responses and endogenous responses to user tastes is important to distniguish between the benefits (reduced toxic engagement due to behavioral responses) and costs (reduced overall engagement with the platform) of the intervention. The behavioral responses depend on user's exposure to socially acceptable content through the algorithm, and can aid policy instruments aimed at reducing engagement with harmful content. However, other endogenous responses, driven by fixed tastes for different types of content, are likely to undermine the effectiveness of such policy measures. Crucially, these responses reduce the time a user spend on the platform, proving costly to the platform in terms of advertising revenues.

5.1.1 Platform and Algorithm

In this model, the algorithm's objective is to serve content according to user's consumption value and the signalling value she derives from sharing posts in her networks. I assume that the structure of user preferences is known to the algorithm and user knows the distribution of content recommendations that the algorithm would make at the beginning of a session.

However, the user can change her view of socially acceptable content through an updating process, only upon viewing the realized sequence of posts that appear in her feed. In this way, exposure through the algorithm distorts user behavior, but the algorithm internalizes this distortion.

The algorithm's objective is to optimize post assignment across various types of content to maximize the total number of posts that a user would view. 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 ν, is endogenously determined by the user for given probability assignments chosen by the algorithm. The algorithm chooses optimally chooses q's to maximize ν for each user, because a user who views a higher number of total posts also consumes a greater number of advertisements.

In optimizing the total number of posts viewed, the algorithm is essentially maximizing the attention paid to various advertisements hosted on the platform. This helps the platform increase the advertisement revenues it generates. Therefore, the firm is interested in maximizing the time users spend on the platform, which is sometimes measured using the total number of posts viewed. I abstract away from the exact process through which this helps the platform optimize revenue from advertisers. 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 and shares, as well as the number of ads shown to each user, and the price charged to advertisers.

5.1.2 User

Consider a social media user who chooses $\mathcal{S}^r \in \{1, \ldots, \nu^r\}$ posts of type-r to engage with (and in particular, to 'share'), where $r \in \{t \text{ (toxic)}, n \text{ (non-toxic)}\}$, and $|\mathcal{S}^r| = S^r$. That is, the user chooses S^r posts to share out of the total ν^r posts of type-r posts viewed. She also picks the utility maximizing number of posts to view, $\nu = \nu^t + \nu^n$, for given assignment probabilities q^t and q^n. Effectively, the user chooses the *proportion* of toxic posts shared as, $s^t = S^t/(S^t + S^n)$.

Choosing s^r generates some *consumption utility* $u(s^r, \nu; q^r, \beta^r, \alpha^r)$, where q^r is the endogenously determined probability with which the algorithm assigns type-r content to a user. Further, β^r is interpreted as users' intrinsic motivation to *view*, and α^r parameterizes users' intrinsic motivation to *share* posts of type r (Bénabou and Tirole, 2006). I assume that $u(s^r, \nu; q^r, \beta^r, \alpha^r)$ is increasing and weakly concave in s^r, and $\frac{du(s^r,\nu;q^r,\beta^r,\alpha^r)}{ds^r}$ is increasing in β^r and α^r. The user's objective is to maximize the utility she derives from consuming posts, as well as the public recognition utility she receives from sharing posts.[29]

For both these channels, I write a model where a user's behavior is reference dependent on the perception of society's engagement behavior, as well as user's own taste toward different types of content. Let $c(s^r; \theta^r, \bar{s}^r)$ be the disutility when signalling action s^r is distinct from average behavior in the population, denoted by $\bar{s}^r$ (Becker, 1991; Besley and Coate, 1992). This is because, the user infers society's tastes for type-r content from the average sharing behavior in the population. I assume that the user wants to conform with this average behavior, by choosing s^r that implies individual tastes close to everyone else's tastes for

[29] consumption utility is obtained from both viewing and sharing posts.

toxic content. In this *action-signalling* model (Butera, Metcalfe, Morrison, and Taubinsky, 2022), users receive value upon choosing action intensity or the probability of sharing type-r content s^r,

$$U^r_{(\text{acts})} = u(s^r, \nu; q^r, \beta^r, \alpha^r) - \nu c(s^r - \theta^r \bar{s}^r)$$

where, θ^r is the weight users put on society's perceived tastes for type-r content, and average shares on a post of type r, $\bar{s}^r$, reveals society's preferences for this type of content as perceived by the particular user.

Users in this model also derive utility from what their action reveals about their true characteristics. They incur a cost if their action signals characteristics that are different from their actual tastes. In other words, the disutility depends on how the user's sharing behavior differs from a reference level which is given by some convex combination of what others do and their own unobserved intrinsic taste. Therefore, the value obtained from sharing posts of type r with probability s^r, upon being assigned q^r proportion of type-r posts by the algorithm is,

$$U^r_{(\text{acts, chars})} = u(s^r, \nu; q^r, \beta^r, \alpha^r) - \nu c(s^r - \theta^r \bar{s}^r - (1-\theta^r)p^r) \tag{12}$$

where, p^r is users' own taste for r-type content, which is not observed by the researcher. When $c(0) = 0$ is the single peak of this 'cost' function, the utility maximization problem is well defined when $c(x) > 0$ for all x. Further, I assume that $\frac{\partial^2 c(s^r - \theta^r \bar{s}^r - (1-\theta^r)p^r)}{(\partial s^r)^2} > 0$, so that the local minima for the $c(\cdot)$ function is achieved at $s^r = \theta^r \bar{s}^r + (1-\theta)p^r$. Then, for the two types of content, expected utility from viewing ν posts in total, and sharing each type of posts with probability s^r is given by

$$U = \nu[u(s^t; q^t, \beta^t, \alpha) + u(s^n; q^n, \beta^n, \alpha)] - \nu c(s^t - \theta \bar{s^t} - (1-\theta)p^t) - \nu c(s^n - \theta \bar{s^n} - (1-\theta)p^n) - \eta \nu^2 \tag{13}$$

where $\eta > 0$ provides the convex cost of increasing total number of posts viewed, which is alternatively interpreted as the cost of spending more time on the platform at the margin. I also assume that the consumption utilities are linearly increasing in the number of posts viewed ν. The key takeaway from this user model is that a user tends to distort the tastes she reveals to the algorithm through her sharing behavior, due to the signalling value she obtains from sharing posts.

5.1.3 Measurement I: $\bar{s}^r \approx q^r$

A user in this framework optimally chooses posts to share according to the consumption value as well as the signalling value. However, she cannot perfectly observe society's tastes $\bar{s}^r$ because engagement with a post is continuously changing. I assume that the perception of society's tastes is inferred from the distribution of the two types of posts that will be viewed, q^r. This is because the algorithm assigns type-r content with probability q^r.

The user infers that other users with similar characteristics also view such content with the same probability because of the personalization algorithm. That is, users are assumed to infer society's tastes for (non-)toxic content using the proportion of each content type that is recommended to them by the algorithm. Then, I rewrite the utility function as

$$U = \nu[u(s^t; q^t, \beta^t, \alpha) + u(s^n; q^n, \beta^n, \alpha)] - \nu c(s^t - \theta q^t - (1-\theta)p^t) - \nu c(s^n - \theta q^n - (1-\theta)p^n) - \eta \nu^2 \tag{14}$$

Therefore, the user's sharing behavior is influenced by exposure to the algorithmically generated content feed, in addition to the intrinsic value they derive from engaging with social media posts. The algorithm, in turn, learns this distortion, and makes content recommendations accordingly. Then, $c(\cdot)$ gives net image pay-off which is assumed to be non-linear (Bénabou and Tirole, 2011).

5.1.4 Timing

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

The platform first optimizes its ad-revenue by algorithmically assigning content that users are more likely to stay and engage with. The algorithms' choice variables are therefore, given by q^t and q^n for each user. However, I assume that the probability with which the algorithm assigns (non-)toxic posts is known to the user. That is, the user is aware of the distribution q^t, but has no way of knowing the posts she will view, a priori.

Next, once the user knows the expected proportion of toxic and non-toxic posts that are assigned to them, she decides when to stop drawing posts from her algorithmically curated feed. In other words, the user decides the total number of posts she will view ν, or the total time she will spend on the platform upon observing the assignment probabilities.

Subsequently, the user sequentially draws posts to view and chooses whether to share each of the viewed posts. The user responds to these assignment probabilities through her engagement behavior, i.e. by sharing a viewed post. This determines s^r. I solve for this general form of the game in Appendix C, but use particular functional forms to estimate relevant parameters derived from the equilibrium conditions in the next section.

5.1.5 Equilibrium

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 others in their network. Users share toxic content so that the signal is close to a convex combination of society's tastes (as perceived by the user), as well as own tastes for toxic content. This structure of user preferences distorts the probabilities with which the algorithm assigns content, away from users' innate tastes for different types of content. This is expected to drive the assignment probabilities into more extreme directions. Users also choose the total number of posts to view, ν.

Consider two time periods in this model, so that $\tau \in \{0, 1\}$ represents the baseline and the intervention period respectively. The algorithm assigns toxic and non-toxic posts to each user i at time τ with probabilities $q^t_{i,\tau}$ and $q^n_{i,\tau}$ respectively. A priori, users know the probability of being assigned (non-)toxic posts, and they use this information to select the optimal number of posts to view, $\nu_{i,\tau}$. Then, a user decides whether to share each post she views with probabilities $(s^t_{i,\tau}, s^n_{i,\tau})$. After sharing a post, user decides whether to draw

another post to view, out of the ranked list of posts assigned by the algorithm. However, the user does not know if the post that is drawn next would be toxic or non-toxic, as she only knows the probability distribution of the algorithm's recommendations.

The solution for this strategic interaction allows me to identify the behavioral responses to recommendations generated using the algorithm. Users are assumed to update their perception of social preferences based on the content they view, and the algorithm in turn internalizes distortions from users' desire to conform. Using backward induction to arrive at the subgame perfect equilibria for the given timing of interactions, first consider the last sub-game where users make sharing decisions to maximize utility. I solve this game using a particular functional form for ease of computation, and introduce user subscripts for each variable at this point. Users maximize utility by choosing the total number of toxic and non-toxic posts to share, i.e. $S^t_{i,\tau}$ and $S^n_{i,\tau}$. Recall that this best response characterizes one of the main outcome variables, i.e. proportion of toxic posts share, $s^t_{i,\tau} = S^t_{i,\tau}/(S^t_{i,\tau} + S^n_{i,\tau})$.

$$\max_{s^t_{i,\tau}, s^n_{i,\tau}} u_{i,\tau} = \nu_{i,\tau}\left[q^t_{i,\tau} + q^n_{i,\tau} - \frac{\left(q^t_{i,\tau}\right)^2}{\beta^t_i} - \frac{\left(q^n_{i,\tau}\right)^2}{\beta^n_i} + \alpha_i(s^t_{i,\tau} + s^n_{i,\tau})\right] - \eta\nu^2_{i,\tau}$$
$$- \nu_{i,\tau}\left(\frac{\alpha_i}{2} + s^t_{i,\tau} - (1-\theta)\left(p^t_i\right)^2 - \theta\left(q^t_{i,\tau}\right)^2\right)^2 - \nu_{i,\tau}\left(\frac{\alpha_i}{2} + s^n_{i,\tau} - (1-\theta)\left(p^n_i\right)^2 - \theta\left(q^n_{i,\tau}\right)^2\right)^2$$

where, recall that p^t_i denotes user i's innate taste for toxic content, and $\nu_{i,\tau}$ is the total number of posts viewed which is endogenously chosen by user i. Then, the utility function captures the pleasure user i derives from viewing $(q^r_{i,\tau} - (q^r_{i,\tau})^2/\beta^r_i)$ and sharing content $(\alpha s^r_{i,\tau})$ in each category. As before, α_i captures the consumption value derived from sharing posts. This utility form also stipulates that users receive some disutility when their sharing behavior deviates from own preferences, as well as perceived societal preference over different types of content. With $\bar{s}^r_{i,\tau}$ measured using $q^r_{i,\tau}$, the following lemma shows that user tastes relative to society's tastes (as perceived through the algorithm) are the main drivers of user behavior.

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

$$s^t_{i,\tau} = (1-\theta)\left(p^t_i\right)^2 + \theta\left(q^t_{i,\tau}\right)^2$$
$$s^n_{i,\tau} = (1-\theta)\left(p^n_i\right)^2 + \theta\left(q^n_{i,\tau}\right)^2$$

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

The optimal sharing strategy is a convex combination of user's own tastes as well as her perception of society's tastes, represented by her algorithmically generated content feed $q^t_{i,\tau}$. The user essentially balances the consumption utility she gets from sharing a post against the disutility of behaving too far from their reference point. The reference point is given as a convex combination of the two variables that constitute the best response, i.e. user's own tastes and perception of society's tastes.

Then, for the given expected number of toxic and non-toxic posts shared, user i solves for ν to maximize total utility from her social media engagement. The algorithm then considers user's endogenous choice of $\nu_{i,\tau}$, to optimally assign $q^t_{i,\tau}$ to maximize user retention on the

platform. User's first order condition yields two solutions for $\nu_{i,\tau}$, one of which is zero and the other is given by the following proporsition.

Proposition 2. *For $q^t_{i,\tau} = 1 - q^n_{i,\tau}$, user chooses total number of posts to view as*

$$\nu_{i,\tau} = \left\{ 0, \frac{1}{2\eta}\left[1+\alpha_i\theta-2\alpha_i q^t_{i,\tau}q^n_{i,\tau}\theta+\alpha_i(1-\theta)\left(\left(p^t_i\right)^2+\left(p^n_i\right)^2\right)-\frac{\left(q^t_{i,\tau}\right)^2}{\beta^t_i}-\frac{\left(q^n_{i,\tau}\right)^2}{\beta^n_i}-\frac{\alpha_i^2}{2}\right]\right\} \quad (15)$$

Proof. Proof in Appendix. □

Note that the non-trivial solution for $\nu_{i,\tau}$ is increasing in user's intrinsic taste for viewing toxic content, β^t_i, but decreasing in the weight user puts on public recognition utility, θ. I later use this result to estimate the effect of the intervention on the total number of posts viewed by users.

Corollary 1. *For user i with $\nu_{i,\tau} > 0$,*

$$\frac{\partial \nu_{i,\tau}}{\partial \theta} \geq 0 \quad (16)$$

That is, the total number of posts viewed by user i is increasing in the weight she puts on public recognition utility.

This shows that the treatment effect is increasing in user's valuation θ. I interpret this result as the fact that users who are more sensitive to the signaling value of sharing posts are more likely to view a greater number of posts on the platform. This also implies that for relatively large assignment probabilities, exogenous changes to the algorithm lead to a larger change the time users spend on the platform, when users are more sensitive to the public recognition utility they derive from sharing posts.

The cross derivative of the total number of posts viewed with respect to the assignment probabilities and the weight users put on public recognition utility is positive. This means that users with larger θ have larger treatment effects on the total number of posts viewed. Next, I endogenize the assignment probabilities by considering the platform's optimization problem, to solve for the customization algorithm. In modeling the algorithm's incentives, this framework highlights the sources of distortion in content allocation. I show that the social signalling aspect of user preferences is likely to increase the assignment of more extreme content to users. In this way, the algorithm affects user behavior by exposiing them to content they can share from, and also by altering their perception of society's tastes.

Using the platform's optimization problem, I show that if users do not derive public recognition utility, the algorithm has less incentive to distort content assignment away from user's innate tastes. The platform's objective is to maximize the total time that users' spend on the platform, because increasing time spent on the platform or the total number of posts also implies that users will view a greater number of advertisements. Therefore, the platform feeds the objective function in (15) to the algorithm, which in turn optimally chooses $q^t_{i,\tau}$ to maximize advertisement revenues.

Proposition 3. *The algorithm assigns toxic posts with probability*

$$q^t_{i,\tau} = \frac{\beta^t_i(\beta^n_i\theta\alpha_i - 1)}{2\alpha_i\beta^t_i\beta^n_i\theta - \beta^t_i - \beta^n_i} \quad (17)$$

Proof. The proof of this proposition flows directly from the first order conditions of an algorithm that is set to maximize $\nu_{i,\tau}$ in (15), by choosing $q^t_{i,\tau}$ optimally. □

The solution for the algorithm's problem provides optimal $q^t_{i,\tau}$, which is increasing in users' tastes for viewing toxic content, β^t_i. While updating perception of social preferences, an optimizing algorithm considers users's innates tastes, as well as distortions due to user perception of society's tastes. The distortions in turn depend on what the algorithm serves the user. It is useful to see that in the absence of this trade-off between own taste and society's tastes, the algorithm sets $q^t_{i,\tau}$ to the relative weight that a user puts on viewing toxic posts. In the absence of behavioral considerations to conform, the algorithm assigns content based on user's intrinsic tastes.

Corollary 2. *In the case of no updating, that is when $\theta = 0$,*

$$q^t_{i,\tau} = \frac{\beta^t_i}{\beta^t_i + \beta^n_i}$$

That is, in the absence of action signalling concerns, the algorithm chooses an assignment probability that is relative to the weight for that content type.

Proof. This claim is obtained by substituting $\theta = 0$ in the assignment probability in (17) in Proposition 3. □

This result illustrates that when there is no distortion on account of users' behavioral updating response to the recommendations, the algorithm learns and caters to users' preferences for viewing toxic content. The distortion is caused by the fact that algorithmic content recommendations impose an *internality* on the user because the recommendation changes users' *perception* of society's preferences for toxic content. In particular, the assignment probabilities demonstrate that the effect of user incentives to signal their conformity with society's norms. To see this, first consider the derivative of the assignment probabilities with respect to θ, which represents the weight users put on the public recognition utility modeled into their preferences.

Lemma 2.

$$\frac{\partial q^t_{i,\tau}}{\partial \theta} = \frac{\beta^t_i \beta^n_i \alpha_i (\beta^t_i - \beta^n_i)}{[2\alpha_i \beta^t_i \beta^n_i \theta - \beta^t_i - \beta^n_i]^2} \gtreqless 0 \tag{18}$$

Proof. This claim is obtained by taking the first derivative of the assignment probability in (17) in Proposition 3, with respect to θ. □

That is, the assignment probabilities for toxic content could be increasing or decreasing in the value a user puts on public recognition utility, depending on the user's innate taste for viewing or consuming toxic content. Then, for given joint distribution of users' consumption values (β^t_i, β^n_i), I show that increasing θ pushes assignment probabilities to the extreme. This depends on user i's initial distribution in the distribution of (β^t_i, β^n_i), because the assignment probabilities only depend on the consumption value when $\theta = 0$ (Lemma 2). I simulate equilibrium assignment probabilities for a randomly generated distribution of user tastes (See Appendix E for simulation details). This shows that the probability of assigning toxic

content increases rapidly with θ for users who like toxic content more than other users. On the other hand, this simulation demonstrates that the algorithm assigns toxic content to users with lower probabilities to users who prefer non-toxic content, for higher values of θ.

The amount of weight that users put on public recognition utilities determines the distance of algorithmic recommendations from their consumption utilities. Furthermore, this distance is increasing in the value of θ or the gains from signalling. These results help in characterizing, and later estimating, the optimal sharing function in terms of the models primitives. In equilibrium, algorithm chooses the probabilities of toxic content for given user preferences as well as the rate of updating perceptions. Users are then able to compute expected number of toxic and non-toxic posts they will share, prior to viewing the realizations of the posts they draw to view, and their respective toxicities.

Proposition 4. *In equilibrium, sharing probabilities are given by user tastes as well as the weight a user puts on their perception of society's preferences.*

$$s^t_{i,\tau} = (1-\theta)p^t_i + \theta\frac{\left(\beta^t_i\right)^2}{(\beta^t_i+\beta^n_i)^2} + O(\theta^2) \tag{19}$$

$$s^n_{i,\tau} = (1-\theta)p^n_i + \theta\frac{\left(\beta^n_i\right)^2}{(\beta^t_i+\beta^n_i)^2} + O(\theta^2) \tag{20}$$

Proof. in Appendix. □

This expresses the sharing function in terms of the model's primitives, which is important for two reasons. First, this is useful in identifying the various channels that determine total platform usage, and the total engagement with problematic content on the platform. In particular, the model identifies the role of users' own tastes and that of aggregate engagement with different types of content. This is important because it explicitly identifies that the importance of signalling conformity, in creating incentives for the algorithm to sort users into different echo chambers.

Second, this framework lays down the foundations for structural estimation that precisely identifies the contribution of user tastes, separately from effect of exposure via recommendation algorithms, in determining engagement with problematic behavior on social media. The next subsection lays out measurement choices made in order to estimate the role exposure via algorithms, separately from the effect of own preferences, in determining engagement with problematic content on social media platforms.

5.1.6 Measurement II: $p^r_i \approx s^r_{i,0}$

The theoretical framework expresses equilibrium behavior in terms of model primitives. However, these primitives may not be observed in the data, and therefore their measurement merits explanation. The data $\mathcal{D}_{i,\tau} = \{S^t_{i,\tau}, S^n_{i,\tau}, V^t_{i,\tau}, V^n_{i,\tau}\}$ 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, $s^r_{i,1} = \frac{S^r_{i,1}}{S^t_{i,1}+S^n_{i,1}}$, for $r \in \{t, n\}$. That is, the proportion of toxic shares out of all the posts shared by user i. I imperfectly measure users' innate preferences for

sharing different types of content with user's baseline behavior ($\tau = 0$), i.e. $\log p_i^t \equiv \log s_{i,0}^t$ and $p_i^n \equiv \log s_{i,0}^n$. This is a reasonable proxy because user behavior at baseline is said to be in equilibrium, so that users do not update their behavior, but continue sharing according to their preferences. The logs are put in place to address skewness in the distribution of toxic shares.

Similarly, I proxy users' relative preference for viewing toxic content with the proportion of toxic posts viewed. This is tenable under the assumption that in equilibrium, the optimizing algorithm assigns toxic content in line with users' tastes for viewing different types of content. Therefore, I set

$$\left(\frac{\beta_i^t}{\beta_i^t + \beta_i^n}\right)^2 \equiv \log \frac{V_{i,1}^t}{V_{i,1}^t + V_{i,1}^n} = \log v_{i,1}^t, \quad \left(\frac{\beta_i^n}{\beta_i^t + \beta_i^n}\right)^2 \equiv \log \frac{V_{i,1}^n}{V_{i,1}^t + V_{i,1}^n} = \log v_{i,1}^n$$

I rewrite the equilibrium sharing probabilities in Lemma 1 as functions of sharing behavior at baseline, and views during the intervention period.

Proposition 5. *The optimal sharing functions are written as*

$$\log s_{i,1}^t = (1-\theta)\log s_{i,0}^t + \theta \log v_{i,1}^t + \mu \log w_i^t \tag{21}$$

$$\log s_{i,1}^n = (1-\theta)\log s_{i,0}^n + \theta \log v_{i,1}^n + \mu \log w_i^n \tag{22}$$

In this representation of the model, w_i^r for all $r \in \{t, n\}$ is the error term from the linearization process above. This is interpreted as the time-invarying taste based shocks at the user level. Proposition 5 gives the required sharing functions that I base the structural estimates off. The parameters of interest cannot be directly estimated through these equations due to certain features of the design of the experiment, which are explained below.

I extend the simulation above to show the optimal sharing function in Appendix E. Assuming users view everything that is assigned to them, the proportion of toxic posts viewed exactly equals the assignment probabilities, i.e. $v_{i,1}^t = q_{i,1}^t$. Then, for different levels of θ, the distortion generated by the social signalling process can produce multiple steady states. The simulation shows that only very high values of θ can produce some unstable steady states. I obtain unique steady states when users do not internalize the signalling costs, or do not care for public recognition utilities.

5.2 Structural Estimation

I use the structural model to measure the rate at which a user updates her sharing behavior to conform with other users in her network. This is because even though the experimental variation identifies user demand by randomly varying allocation of toxic content in the treatment group, the mechanisms that drive user responses are not identified. I estimate the model in order to identify the role of users innate preferences, separately from the rate at which users update their behavior in response to changes in the algorithmic recommendations. I postulate that users change their behavior is response to the treatment as they infer society's tastes for differernt types of content using the algorithm's recommendations. Accordingly, users update their sharing behavior, at some rate θ, to remain in step with other similar users.

While the platform employs it's personalization algorithm to cater user incentives outlined above, the modeled algorithm is a fairly coarse representation of more sophisticated technology. My objective is to mimic the incentives of a simple algorithm in order to analyze user responses, and not to analyze the algorithm itself. This model shows that the algorithm caters to users' innate tastes when users do not care as much for public recognition utility. On the other hand, the algorithm's recommendations are distroted in more extreme directions when users benefit from the social image they project by sharing posts.

5.2.1 Intervention

Since exposure and engagement with toxic content are endogenous variables due to the recommendation algorithm, the demand and supply factors driving problematic behavior cannot be disentangled in the control group. This is the main identification problem that the experiment solves. This is because the algorithm is trained on some underlying user preferences that also determine user behavior, but are not observed by the researcher. 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: effect of *exposure* to toxic content that is usually recommended by the algorithm, 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 preference for conformity. This is because a draw of posts is picked randomly for the treated users each day, during the intervention period.

The Central Limit Theorem predicts that under a daily random draw of assignment probabilities, the probability of assigning toxic posts for the treated group follows a normal distribution centred 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 $\sqrt{|\{i|D_i = 1\}|}$). In fact, exposure to a particular type of content is almost constant among treated users, rendering the relationship between exposure and engagement *unidentifiable* in the treatment group.

I further belabour this point about the absence of a straight forward identification strategy. 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

$$\mathrm{E}\left[\log\frac{s^t_{i,1}}{s^t_{i,0}}\middle|D_i=1\right]-\mathrm{E}\left[\log\frac{s^t_{i,1}}{s^t_{i,0}}\middle|D_i=0\right]=\theta\left(\mathrm{E}\left[\log\frac{v^t_{i,1}}{v^t_{i,0}}\middle|D_i=1\right]-\mathrm{E}\left[\log\frac{v^t_{i,1}}{v^t_{i,0}}\middle|D_i=0\right]\right)$$

This does not identify the main parameter of interest θ, because the algorithm and control users continue to remain in steady state during the intervention period. As a result, differences in toxic views from baseline are seen as random deviations from the steady state. I exploit unique features of this setting to model behavior and estimate pertinent parameters that are not be estimable using a standard difference-in-difference approach.

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 in comparisons between treatment and control users, 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 from problematic content (measured by baseline shares), **(2)** users' perception of society's preferences (measured by assignment probabilities).

Let D_i be a binary variable denoting treatment status, so that $D_i = 0$ when i was assigned to the control group. The intervention described in Section 4 is administered during $\tau = 1$, and I consider the first month of implementing this intervention (February 10, 2023 to March 10, 2023) for my empirical analysis. By design, treated users were assigned toxic content with probability that is drawn at random from the control distribution of $q^t_{i,1}$, each day during the first month of the intervention.

On average, treated users were assigned toxic content with probability $\bar{q^t} = E[q^t_{i,1}|D_i = 0]$ by design. Analogously, on average, treated users were assigned non-toxic content with probability $\bar{q^n} = E[q^n_{i,1}|D_i = 0]$. This is because of the CLT, as the assignment probabilities are independently resampled each day, for treated users from the control distribution of $q_{i,1}$ during the intervention period. Control users continue to remain in their steady state, so that these users view toxic and non-toxic content according to the algorithmic assignment in equilibrium.

5.2.2 Sharing Behavior

According to Proposition 5, users update their sharing behavior using new posts of the type they have viewed by some parameter θ. This is because the algorithm's content recommendation allow the user to infer the tastes of other similar users, and users update their sharing behavior to conform with these inferred tastes. Considering the exponential transformation of (21), the probability of sharing toxic or non-toxic content, during the intervention period ($\tau = 1$) is,

$$s^t_{i,1}(D_i) = \big(v^t_{i,1}(q^t_{i,1}, D_i)\big)^{\theta}\big(w^t_i\big)^{\mu}\big(s^t_{i,0}\big)^{1-\theta} \tag{23}$$

$$s^n_{i,1}(D_i) = \big(v^n_{i,1}(q^n_{i,1}, D_i)\big)^{\theta}\big(w^n_i\big)^{\mu}\big(s^n_{i,0}\big)^{1-\theta} \tag{24}$$

where, $v^t_{i,1}(q^t_{i,1}, D_i)$ and $v^n_{i,1}(q^n_{i,1}, D_i)$ are the proportions of toxic and non-toxic content viewed for given treatment assignment (D_i), respectively. Then, the probability of sharing toxic content, $s^t_{i,1}(D_i)$ for given treatment assignment describes user behavior during the intervention period. This dynamic model specifies sharing behavior with respect to toxic posts, during the intervention period, as a function of the proportion of toxic content viewed in the same time period, time in-varying heterogeneity in user preferences, and user behavior at baseline.

The reduced form estimates indicated that user behavior is not very malleable in the face of the administered intervention. This is because even though treated users saw a lower proportion of toxic posts, I find a smaller reduction in the proportion of toxic posts shared among the treated. This is because exposure to new information changes user behavior to a limited degree. Then, a user incurs costs from sharing posts that are not aligned with her true type.

5.2.3 Identifying Conditions

This model of interaction between recommendation algorithms and social media users is said to be in steady state when the probabilities of assigning toxic content and non-toxic content ($[q^t_{i,\tau}(D_i), q^n_{i,\tau}(D_i)]$), as well as the probabilities of sharing toxic and non-toxic content ($[s^t_{i,\tau}(D_i), s^n_{i,\tau}(D_i)]$) are stable over time. This steady state condition is the main 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, allowing for the identification of the updating parameter θ, which does not change over time or by treatment status. Therefore, users' sharing behavior in steady state is given by the conditions

$$s^t_{i,0}\big(v^t_{i,0}, w^t_i\big) = s^t_{i,1}\big(v^t_{i,1}(q^t_{i,1}, D_i), s^t_{i,0}, w^t_i, D_i\big) \tag{25}$$

$$s^n_{i,0}\big(v^n_{i,0}, w^n_i\big) = s^n_{i,1}\big(v^n_{i,1}(q^n_{i,1}, D_i), s^n_{i,0}, w^n_i, D_i\big) \tag{26}$$

when the intervention exogenously changes probabilities of assigning content of different types. At $\tau = 1$, users in the control group continue to remain in the steady state, which means that θ cannot be correctly estimated using data on shares and views for control users. This is because the proportion of different types of content viewed by control users, is correlated with sharing behavior at baseline, on account of the assignment probabilities generated by the algorithm. Recall that the control algorithm and the outcome variable are affected by unobserved user heterogeneity, which is not observed by the researcher. Therefore, sharing and viewing are endogenously determined for control users.

Lemma 3. *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 biased.*

In other words, for control users, the probability of viewing toxic $v^t_{i,1}(q^t_{i,1}, D_i = 0)$ depends on baseline sharing behavior (which represent users innate tastes for toxic content, p^r_i), along with the probability of sharing toxic content during the intervention period. This is because control users continue to remain in steady state during the intervention period. As a result, estimates of θ generated from the control sample would be biased.

For user i in the treatment group, I rewrite the probability of sharing content. I maintain the assumption that sharing behavior is endogenously determined by baseline behavioral patterns, as well as exposure to new content. Then, the conditional probabilities of sharing toxic and non-toxic content provide the main identifying assumptions because endogenous sharing behavior at $\tau = 1$ or $\tau = 0$ is *a priori* expected to be equal.

Lemma 4. *θ is identified when the following assumptions are satisfied,*

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

$$s^n_{i,0}(D_i = 1) = \big(v^n_{i,1}(\bar{q^n}, D_i = 1)\big)^{\theta}\big(w^n_i\big)^{\mu}\big(s^n_{i,0}\big)^{1-\theta} \tag{A3}$$

This is because, there is no reason to believe that the sharing behavior of any user i would be different in the two time periods in the absence of any exogenous changes in assignment probabilities. I test this assumption in the discussion of the model in Section 5.3. Given these identifying assumptions, θ can be estimated using data on current and past shares because $\bar{q^t} \perp s^t_{i,0}$ and $\bar{q^n} \perp s^n_{i,0}$ by design of the experiment. This implies that the right hand side terms in the expressions above are not correlated.

As explained in Section 5.2.1, I do not observe enough variation is toxic exposure for treated users during the intervention period. However, the difference in assignment probabilities between baseline and intervention period provides heterogeneity in treatment intensity which is exogenous to user tastes.

The effect on difference in sharing behavior over time is correctly estimated using the difference in exposure to toxic content over time, as this accounts for unobserved heterogeneity in sharing behavior. Since, exposure to toxic content among the treated during the intervention period is constant on average, the estimates from this differencing strategy effectively only uses variation in baseline exposure among toxic users. 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.

I estimate θ using the relationship between baseline views and current shares among treated users, which is identified using the steady state condition in (A2). Then, this provides an estimation strategy that is independent of user's preference shocks which are not observed by either the algorithm, or the researcher.

Proposition 6. *For some updating parameter θ and constant share of toxic posts λ, 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,*

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

Proof. In appendix. □

Intuitively, the constructed outcome in Proposition 6 accounts for unobserved user level differences in preference for sharing any type of content. Proposition 6 effectively provides the estimating equation from the relationship between toxic shares and toxic views for treated users. Next, I specify the estimation strategy as well as the assumptions required to implement it.

5.2.4 Estimation

I estimate the updating parameter θ from the relationship between baseline exposure to toxic content ($v^t_{i,0}(1)$), and the odds of sharing toxic content during the intervention period ($s^t_{i,1}(1)/s^n_{i,1}(1)$) among *treated* users only. This helps in estimating the 'stickiness' of the behavioral effect of changing content recommendations made by an algorithm. θ can be 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. As a result, the behavioral responses are identified using the relationship between odds of sharing toxic content prior and during the intervention period.

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). Therefore, I employ a first differences approach to estimate θ using linear approximations of the log ratios. 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 7. *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

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

(SA2) The behavioral effect is constant across time and users.

(SA3) users engagement behavior 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, $\mathrm{E}[q^t_{i,1}(D_i)|w^t_i, s^t_{i,0}, D_i] = \mathrm{E}[q^t_{i,1}(D_i)|s^t_{i,0}, D_i]$

Then, θ can be estimated using the following regression equation

$$\mathrm{E}[\Delta B_{i,1}|D_i = 1] = \gamma_0 + \gamma_1 A_{i,0} \tag{27}$$

where, $\gamma_1 = -\theta$.

Proof. In Appendix. □

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. 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.

Estimating θ allows me to distinguish between the two channels that drive the treatment effect of switching off the content recommendation algorithms. These are users' innate tastes from problematic content (measured by baseline shares), and users' perception of society's preferences (measured by assignment probabilities). This model provides the machinery to identify estimates of the stickiness in human behavior, in addition to the correct interpretation of these results. Furthermore, this interpretation is generalizeable 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 platform that is in steady state. I exploit unique features of this setting to estimate parameters that may not be estimable using a standard difference-in-difference approach.

5.2.5 Measurement Error

The estimation strategy above shows that θ is measured using a sample of users in the treatment group, whose exposure to content was completely random. θ cannot be estimated using a difference-in-difference strategy because both views and shares, for each type of content, are in steady state for control users. 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, this measurement error can be represented as classical linear measurement error (Schennach, 2016).

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.

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

where, $\Delta v^t_{i,0}$ and $\Delta v^n_{i,0}$ denote the measurement error in the proportion of toxic and non-toxic content viewed respectively. In general, assume that $Cov(v^{t*}_{i,0}(1), \Delta v^t_{i,0}(1)) = 0$ and $Cov(v^{n*}_{i,0}(1), \Delta v^n_{i,0}(1)) = 0$. The estimators of θ 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 subsequently referred to as $half2$). However, assuming that the measurement errors pertaining to each half of the posts, per user, are uncorrelated to each other, the fraction over the first half of posts can be instrumented by the second half of the posts. That is to say,

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

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

Proposition 8. *Instrumenting average toxic views for each user with the average of one half of posts viewed provides unbiased estimates.*

Proof. In Appendix. □

The OLS estimates suggest that this strategy does not account for measurement error, indeed 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. Crucially, users do not seem to change their behavior upon seeing content that other similar users are engaging with.

5.3 Discussion

I have developed a framework to highlight the channels that lead to user exposure to and engagement with harmful social media content. The algorithm is programmed to maximize the time users spend on the platform, by optimally choosing the content to recommend to each user. This enables the platform to generate revenues from advertising, and so shutting down the algorithm can be costly for the firms. On the other hand, users decide the proportion of different types of posts to share with their friends, as well as the total time they spend on the platform. User behavior is assumed to be determined by a combination of her own tastes as well as her perception of social preferences for different types of content. That is, the algorithm not only provides users with toxic content for her to share from, but in exposing users' to toxic content, the algorithm also alters their perception of society's preferences with respect to toxic content (DellaVigna and Gentzkow, 2010).

The structural model identifies the role of the algorithm in helping users update their behavior in line with the society's tastes for different kinds of content. This is important because the user obtains value from public recognition of conformity with society. As a result, the model predicts the malleability of user behavior by estimating θ. This is because user's innate tastes are considered to be unchanged, as is standard in behavioral models of social interactions (Akerlof and Kranton, 2000). However, the model makes a set of assumptions that help in identifying the parameters of interest, and simplifying a convulated setting. These assumptions need to be validated.

5.3.1 Simplifying Assumptions

In writing the utility for structural estimation, I made five simplifying assumptions: **(1)** consumption as well as signalling utility is additively separable for each content type, **(2)** the consumption utility from *sharing* is equal for both types of content, i.e. $\alpha^t = \alpha^n = \alpha$, **(3)** action-signalling utility from *sharing* is also equal for both types of content, i.e. $\theta^t = \theta^n = \theta$, **(4)** user behavior in the action-signalling model is updated at some constant rate θ across all users, **(5)** deviating from the reference point of own and society's tastes generates disutility which is quadratic in nature.

The most general form of the model presented above is restrictive in the sense that consumer utility is additive over the two types of posts. This 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. Additionally, the path dependency of user behavior on all the distribution of all types of posts is modeled through the action-signalling utility separately.

Out of these simplifying assumptions, I am able to test a combination of the second and third assumptions. That is, I can observe whether the consumption and signalling value from sharing toxic and non-toxic content is equal. This is validated in Table 5, as I test for the equality of coefficients, stacking two 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 constant effects assumption with respect to the rate of updating user behavior in the action-signalling model. Figure A.8 supports this assumption, as the estimates of θ obtained from samples of different 'types' of users are indinstinguishable from each other. I reject the hypothesis that the rate of updating user behavior is different across different types of users using Table B.3.

Finally, I have assumed the costs of using social media to be quadratic for ease of computation. In doing so, 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 (Butera et al., 2022).

5.3.2 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 continue to remain in steady state during the intervention and were chosen randomly, I test this assumption in the sample of control users,

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

In order to test this assumption, I estimate parameters of the following regression,

$$s^t_{i,1}(D_i = 0) = \delta_0 + \delta_1 s^t_{i,0}(D_i = 0) + \varepsilon_{i,1}$$

Under assumption (IA) I expect $\delta_0 = 0$ and $\delta_1 = 1$ in the sample of control users. As before, I correct for measurement error in computing the proportion of toxic posts shared, and also normalize each variable to have zero mean. Table 6 shows that the hypothesized values of δ_0 and δ_1 cannot be rejected, in the measurement error corrected case.

6 Structural Estimates

A difference-in-difference estimator for elasticity of toxic shares with respect to toxic views is not estimable among control users because the proportion of toxic views is in steady state. I estimate the structural relationship between the differences in probability of sharing toxic

content over time, and the proportion of toxic posts viewed during the intervention period using the following specification,

$$\Delta B_{i,1}(D_i = 1) = \gamma_0 + \gamma_1 A_{i,0}(D_i = 1) + \varepsilon_{i,1} \tag{28}$$

where, $A_{i,0}(D_i = 1)$ is the difference between the proportions of toxic and non-toxic content viewed at baseline and $\varepsilon_{i,1}$ is assumed to be a random error term. This exercise is performed on treated users only, and it estimates the share of toxic engagement that is driven by user behavior at baseline via feed-ranking algorithms. The resultant estimates are correctly identified because the treated users are exposed to toxic content during the intervention period purely by chance, therefore eliminating endogeneity in views and shares. Further, even though γ_1 can be estimated in the control group, the estimates are likely to be biased because toxic views and toxic shares in this case depend on toxic shares at baseline, because of the content recommendation algorithm.

It has been shown that if the relationship of interest was indeed linear, then $\hat{\gamma_1} = -\hat{\theta}$. However, I first assess if the relationship between these two variables is linear in Figure A.9a. Due the skewness of the distribution of toxic views at baseline, I flexibly fit the probability of sharing toxic content with respect to the difference proportion of toxic posts and non-toxic posts viewed at baseline (as an approximation to the log-ratios).

The lowess curve in Figure A.9a approximates a linear relationship between difference in past exposure and the difference in the difference over time for current engagement with toxic content, in the treatment group. I also show the fitted values from a linear specification. This shows the discrepancies between linear and more non-parametric approaches are small.

6.1 OLS Estimates

Table **??** shows that for $\hat{\gamma_1} = -\hat{\theta}$, or the estimated effect of a 1% decrease in the proportion of toxic posts viewed during the intervention period decreases the change in proportion of toxic posts shared by $-\hat{\theta}\% = 0.08\%$. This also demonstrates that estimated stickiness in user behavior, or the elasticity in sharing behavior with respect to baseline sharing behavior, $(1 - \hat{\theta})$ is close to 0.92. Therefore, upon being exposed to toxic content randomly, users did not fully update their behavior in line with content they saw during the intervention period.

These OLS estimates however, are likely to be biased downwards. This is because the data on proportion of toxic or non-toxic posts viewed is likely to suffer from measurement error, due to the fact that users sample posts to view after the algorithm assigns them. I use the IV estimation strategy outlined in the previous section, to correct for this possibility.

6.2 IV Estimates

The IV estimates, in Column (2) of Table 7, indicate that the measurement error indeed led to attenuation bias in the main OLS estimates shown in column (1) of Table **??**. The first column in Table 7 shows the strength of the first stage for the IV estimate. 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). This is the most accurate estimate of θ that can be produced by this framework. The measurement

error corrected estimate shows that a 1% reduction in exposure to toxic content reduces engagement with toxic content by 0.16%.

This has 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. Note that reductions in exposure to toxic content does not have a one-to-one relationship with change in sharing behavior. Table 7 instead, shows that user behavior on social media platforms depends on pre-existing user behavior by a significant degree.

These results are consistent with a model of user behavior where user response is determined by both current and previous exposure to toxic content. 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. Therefore, while users largely follow their old behavioral patterns, user behavior is somewhat malleable according to new exposure to toxic content.

6.3 Validation

I validate my estimation procedure that measures the rate at which users update their sharing behavior upon being randomly exposed to more non-toxic content during the intervention period. This model correctly estimates the updating-behavior only for treated users because exposure to toxic content in the baseline period is related to the engagement with such content during the intervention period 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) Indirect:** Feed-ranking algorithms that expose users to toxic content to maximize engagement using prior user behavior. Since both these channels are correlated by design of the algorithm, the said relationship in the control group is shown to suffer from omitted variable bias.

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 $(v^t_{i,0}/v^n_{i,0})$ the main estimating equation (43). This relationship may not be 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. This is because control users are in steady state, so the ratio of shares at baseline and during the intervention period is expected to be equal.

The estimates are therefore, expected to be distinct from the main estimates above (See Table B.5 and Figure A.10a). This validates the main estimation strategy because the estimates from the same exercises among samples of treated and control users yields very distinct results. Additionally, Figure A.9b shows that exposure to toxic content *during the intervention period* has a much smaller effect on the odds of sharing such content.

7 Conclusion

Recommendation algorithms determine the substance and ranking of content that social media users view on various platforms, including Netflix, Facebook, TikTok and Google Search Ads. This is concerning especially if the algorithm exposes users to misinformation, or hate speech, as these aggregation technologies can lead users into echo chambers. This is thought to increase political polarization across the world. However, less attention is paid to the role of humans in shaping digital spaces like social media platforms, where more than half of American teenagers are known to spend more than four hours a day.[30]

This paper studies the role of algorithmically generated content feeds as well as subsequent behavioral responses in driving online engagement with different types of content. I conducted a large-scale randomized evaluation of such algorithms by effectively turning it off for two 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 behavior and therefore, distinguishes the influence of technology from that of human behavior. 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, the reduction in the odds of sharing toxic content were lower. I then developed a structural model which shows that this may be the case because human behavior at baseline drives more than three-quarters of the variation in odds of sharing toxic content. This means that users update their behavioral patterns in line with exposure to new (mis)information only partially, because 1% decrease in proportion of toxic posts viewed at baseline corresponds to a 0.84% decrease in the odds of sharing toxic content.

These structural estimates are consistent with a user model where engagement adds to users' consumption value and enables them to receive public recognition utility by signalling their action and characteristics. In this model, users are endowed with some inherent taste for toxic content, which is measured by their baseline sharing behavior. Additionally, users infer society's tastes for different types of content from the algorithmically generated content feed. This is because users know that the algorithm recommends similar content to users with similar characteristics.

My findings have important implications for policy makers 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.

[30] https://news.gallup.com/poll/512576/teens-spend-average-hours-social-media-per-day.aspx

References

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.

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

S. N. Ali and R. Bénabou. Image versus information: Changing societal norms and optimal privacy. *American Economic Journal: Microeconomics*, 12(3):116–164, 2020.

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

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

G. Aridor. Drivers of digital attention: Evidence from a social media experiment. 2022.

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

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

D. Arnold, W. Dobbie, and P. Hull. Measuring racial discrimination in algorithms. In *AEA Papers and Proceedings*, volume 111, pages 49–54. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2021.

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

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.

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

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

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

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

R. Bénabou and J. Tirole. Identity, morals, and taboos: Beliefs as assets. *The quarterly journal of economics*, 126(2):805–855, 2011.

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

M. Beraja, A. Kao, D. Y. Yang, and N. Yuchtman. *Exporting the surveillance state via trade in AI*. Brookings Institution, 2023b.

M. Beraja, D. Y. Yang, and N. Yuchtman. Data-intensive innovation and the state: evidence from ai firms in china. *The Review of Economic Studies*, 90(4):1701–1723, 2023c.

T. Besley and S. Coate. Workfare versus welfare: Incentive arguments for work requirements in poverty-alleviation programs. *The American Economic Review*, 82(1):249–261, 1992.

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. Boxell, M. Gentzkow, and J. M. Shapiro. Cross-country trends in affective polarization. *Review of Economics and Statistics*, pages 1–60, 2022.

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

L. Bursztyn and R. Jensen. Social image and economic behavior in the field: Identifying, understanding, and shaping social pressure. *Annual Review of Economics*, 9:131–153, 2017.

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, 2020a.

L. Bursztyn, I. K. Haaland, A. Rao, and C. P. Roth. Disguising prejudice: Popular rationales as excuses for intolerant expression. Technical report, National Bureau of Economic Research, 2020b.

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.

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.

S. DellaVigna and M. Gentzkow. Persuasion: empirical evidence. *Annu. Rev. Econ.*, 2(1): 643–669, 2010.

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 and E. La Ferrara. Economic and social impacts of the media. In *Handbook of media economics*, volume 1, pages 723–768. Elsevier, 2015.

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.

R. Durante and B. Knight. Partisan control, media bias, and viewer responses: Evidence from berlusconi's italy. *Journal of the European Economic Association*, 10(3):451–481, 2012.

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.

K. Garimella and D. Eckles. Image based misinformation on whatsapp. In *Proceedings of the Thirteenth International AAAI Conference on Web and Social Media (ICWSM 2019)*, 2017.

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

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

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

G. H. Golub and C. Reinsch. Singular value decomposition and least squares solutions. In *Handbook for Automatic Computation: Volume II: Linear Algebra*, pages 134–151. Springer, 1971.

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.

Y. Halberstam and B. Knight. Homophily, group size, and the diffusion of political information in social networks: Evidence from twitter. *Journal of public economics*, 143:73–88, 2016.

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.

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

R. Jensen and E. Oster. The power of tv: Cable television and women's status in india. *The Quarterly Journal of Economics*, 124(3):1057–1094, 2009.

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

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

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

J. Kleinberg, J. Ludwig, S. Mullainathan, and A. Rambachan. Algorithmic fairness. In *Aea papers and proceedings*, volume 108, pages 22–27. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2018.

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.

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. Fragile algorithms and fallible decision-makers: lessons from the justice system. *Journal of Economic Perspectives*, 35(4):71–96, 2021.

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

M. Mishra, P. Yan, and R. Schroeder. Tiktok politics: Tit for tat on the india–china cyberspace frontier. *International Journal of Communication*, 16:26, 2022.

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

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.

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

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

Statista. We are social datareportal hootsuite, 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/`.

C. R. Sunstein. *Republic. com*. Princeton university press, 2001.

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.

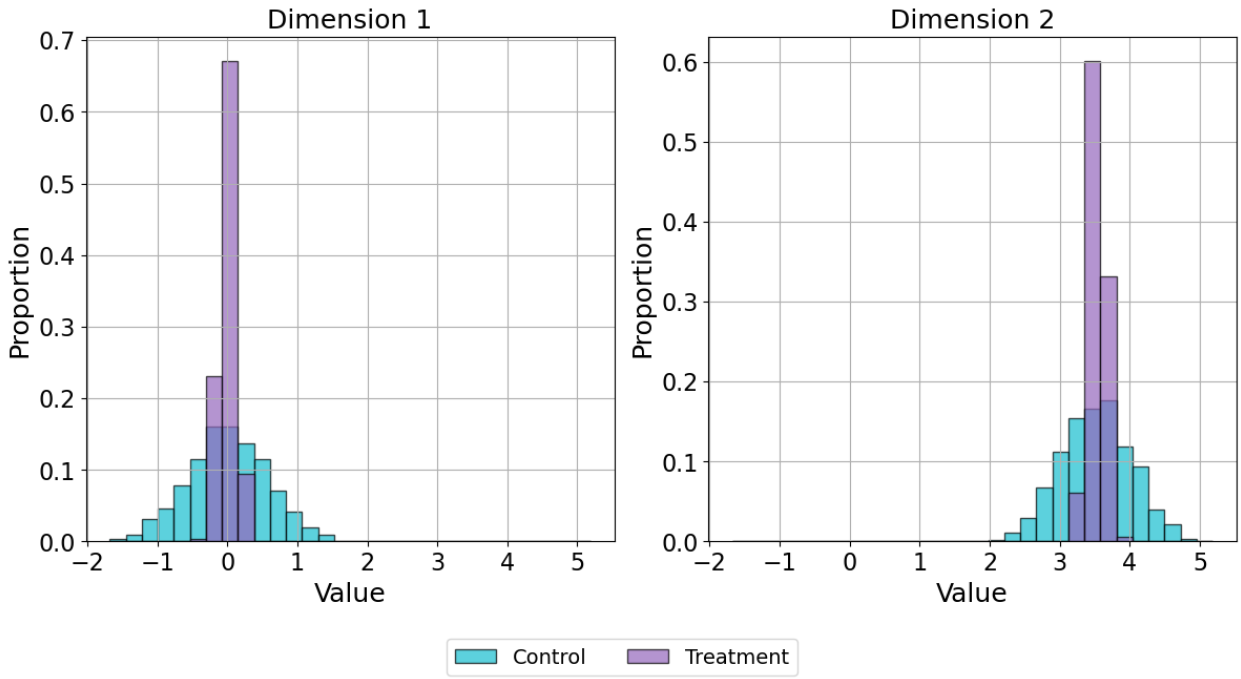
J. J. Van Bavel, S. Rathje, E. Harris, C. Robertson, and A. Sternisko. How social media shapes polarization. *Trends in Cognitive Sciences*, 25(11):913–916, 2021.

M. E. Wall, A. Rechtsteiner, and L. M. Rocha. Singular value decomposition and principal component analysis. In *A practical approach to microarray data analysis*, pages 91–109. Springer, 2003.

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

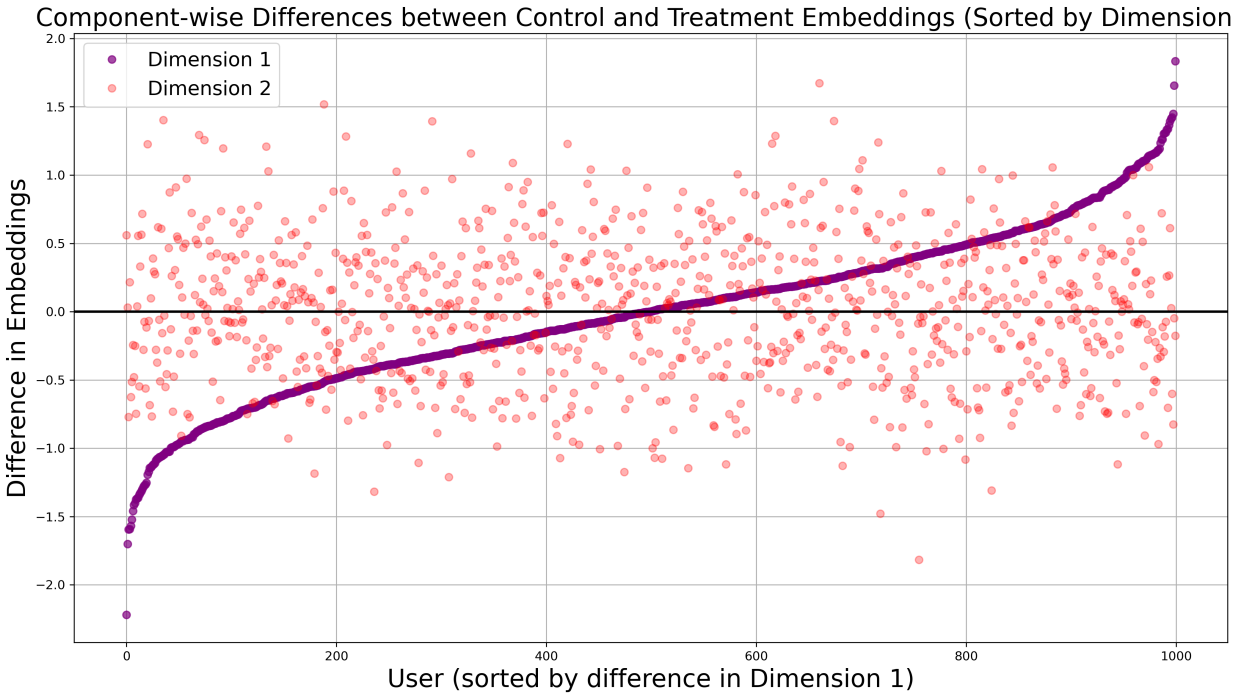
Tables and Figures

Figure 1: Distribution of simulated two-dimensional embedding vectors

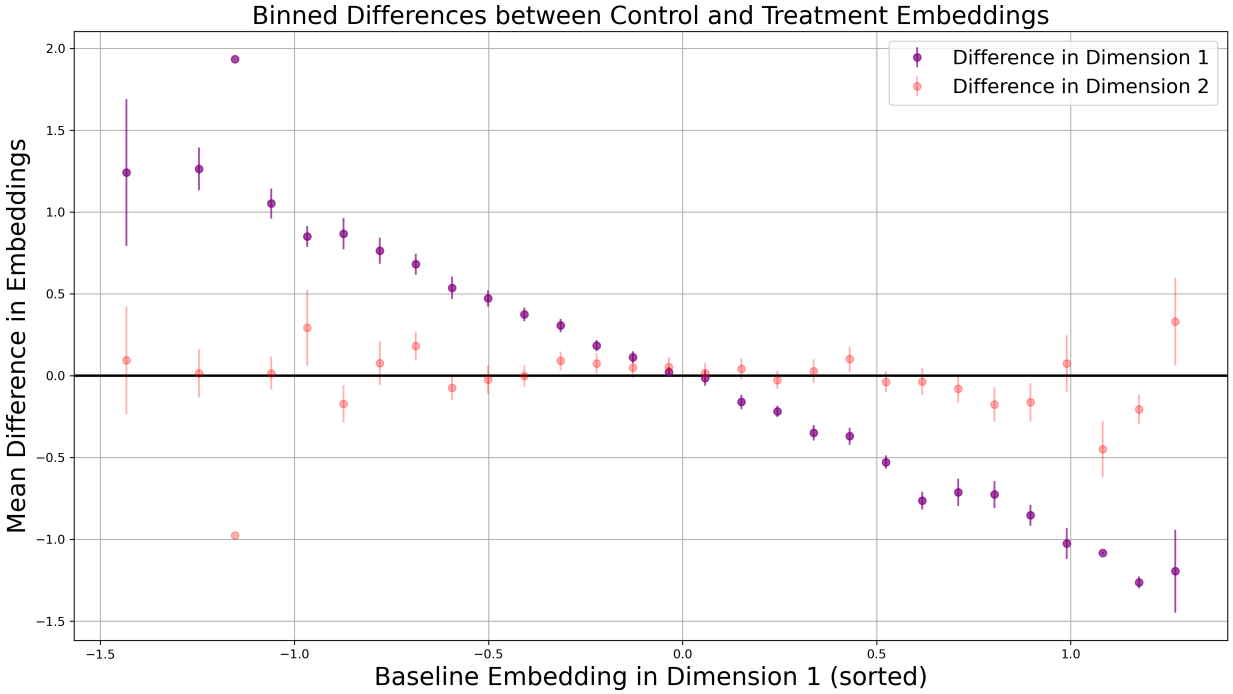


Notes: This graph shows that the two dimensions (components) of the simulated embedding vector follow a Gaussian distribution centred around the mean of each embedding dimension. The spread of control user embeddings is larger than the embeddings generated for treated users. This is because the treatment embeddings are drawn each day from a given sample of control embeddings during the intervention period (Central Limit Theorem). The algorithm used to simulate the embeddings for both treatment and control groups uses Singular Value Decomposition to factorize the matrix of engagement in the control group. Embeddings from the treatment group were uniformly drawn from an epsilon ball centred around the mean control embedding. Details of the simulation are contained in Appendix E.

Figure 2: Difference between embeddings assigned during intervention period and baseline



(a) Raw distribution of difference in embeddings



(b) Mean difference in embeddings, sorted by embedding values at baseline

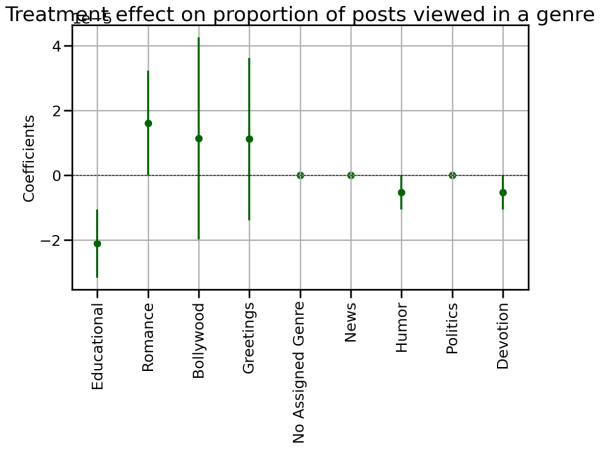
Notes: This graph shows that the treatment has a smaller effect on users with embeddings closer to the average at the baseline. On the other hand, users assigned more extreme embeddings at baseline had bigger absolute effects in content exposure. The algorithm used to simulate the embeddings for both treatment and control groups uses Singular Value Decomposition to factorize the matrix of engagement in the control group. Embeddings from the treatment group were uniformly drawn from an epsilon ball centred around the mean control embedding. Details of the simulation are contained in Appendix E.

Table 1: Balance in treatment assignment across all observable characteristics

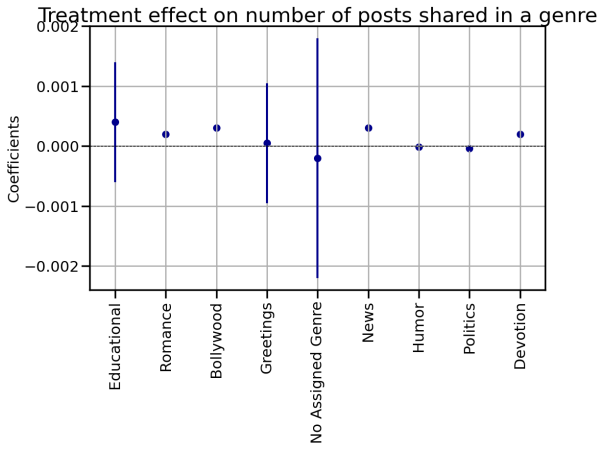
Variable	Mean	Coef.	Std.Err.
Observable User Characteristics			
State: noState	0.631	-0.029	0.084
City: bangalore urban	0.001	-0.006	0.028
City: bhilwara	0.001	0.027	0.034
City: dehradun	0.001	0.018	0.028
City: hanumangarh	0.001	0.013	0.031
City: haridwar	0.002	-0.005	0.025
City: jhansi	0.001	0.016	0.030
City: jhunjhunu	0.002	-0.001	0.026
City: khandwa	0.001	0.004	0.035
City: kolhapur	0.001	-0.044	0.032
City: patna	0.003	-0.009	0.023
City: sant ravidas nagar	0.001	0.027	0.034
City: sri ganganagar	0.001	-0.001	0.030
City: sultanpur	0.001	0.004	0.026
City: valsad	0.001	0.032	0.033
Gender: M	0.695	-0.001	0.003
Week: 2016-04	0.000	0.028	1.180
Week: 2016-29	0.000	0.816	1.201
Week: 2019-17	0.001	0.271	1.181
Week: 2021-37	0.004	0.300	1.180
Baseline Behavior			
Average Toxic Posts Viewed	2.681	0.000	0.001
Average Toxic Posts Viewed (%)	36.890	-0.000	0.000
Number of Posts Shared (%)	761.982	0.000	0.000
Number of Posts Shared	21.411	-0.000	0.000
Number of Logins	20.499	0.000	0.000
N	251704		
F-statistic:	0.887	P-value:	0.689

Notes: This table shows the results of a single regression using all observable characteristics, run at the user level. Coefficients for a randomly selected set of observable characteristics are shown in the table. $D_i = \beta_0 + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where D_i is binary variable taking value 1, when user i was assigned to the treatment group. Weeks correspond to the date on which a user created her account. None of the observable characteristics are correlated with treatment assignment. Cannot reject null hypothesis of joint insignificance. Standard errors are robust at user level.

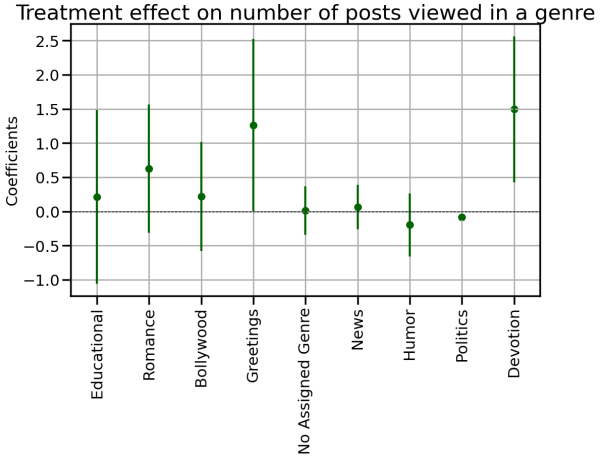
Figure 3: Balance in viewing and sharing content from various genres at baseline



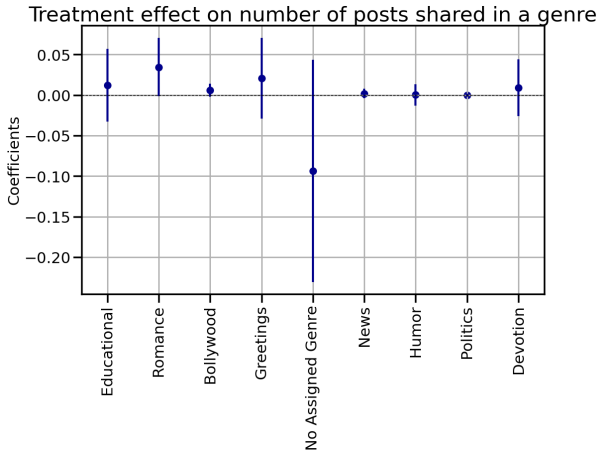
(a) Proportion posts viewed. F-stat: 0.809, p-value: 0.608



(b) Proportion posts shared. F-stat: 0.798, p-value: 0.691



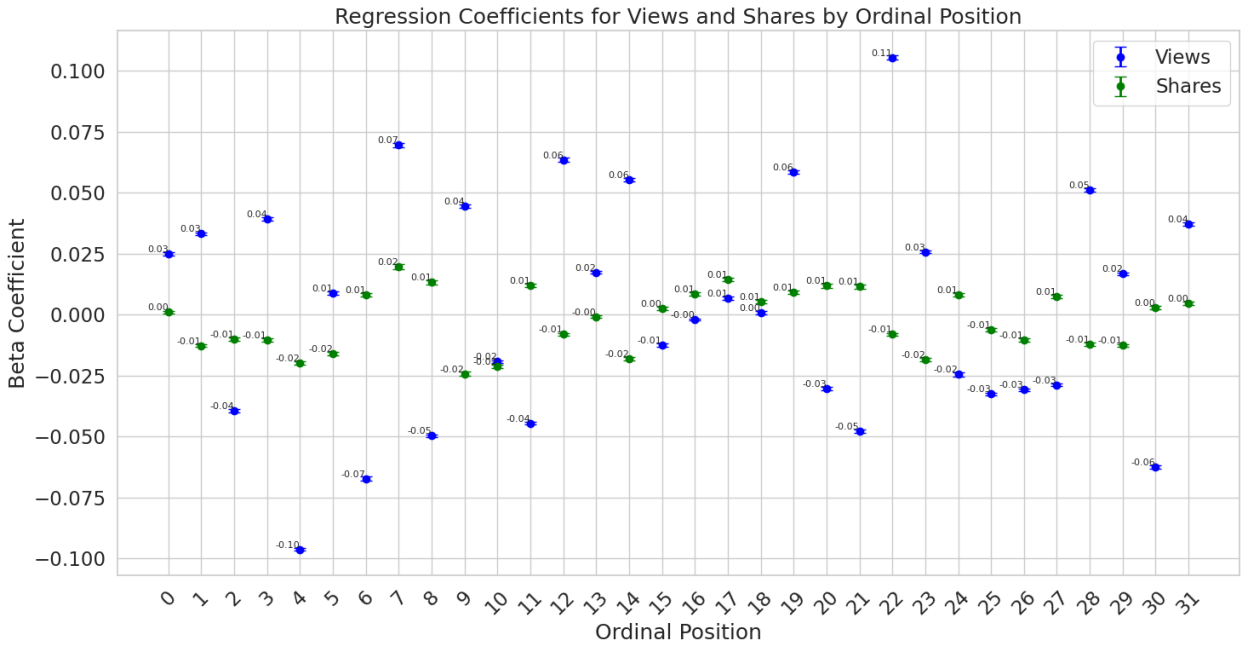
(c) Number of posts viewed. F-stat: 0.828, p-value: 0.590



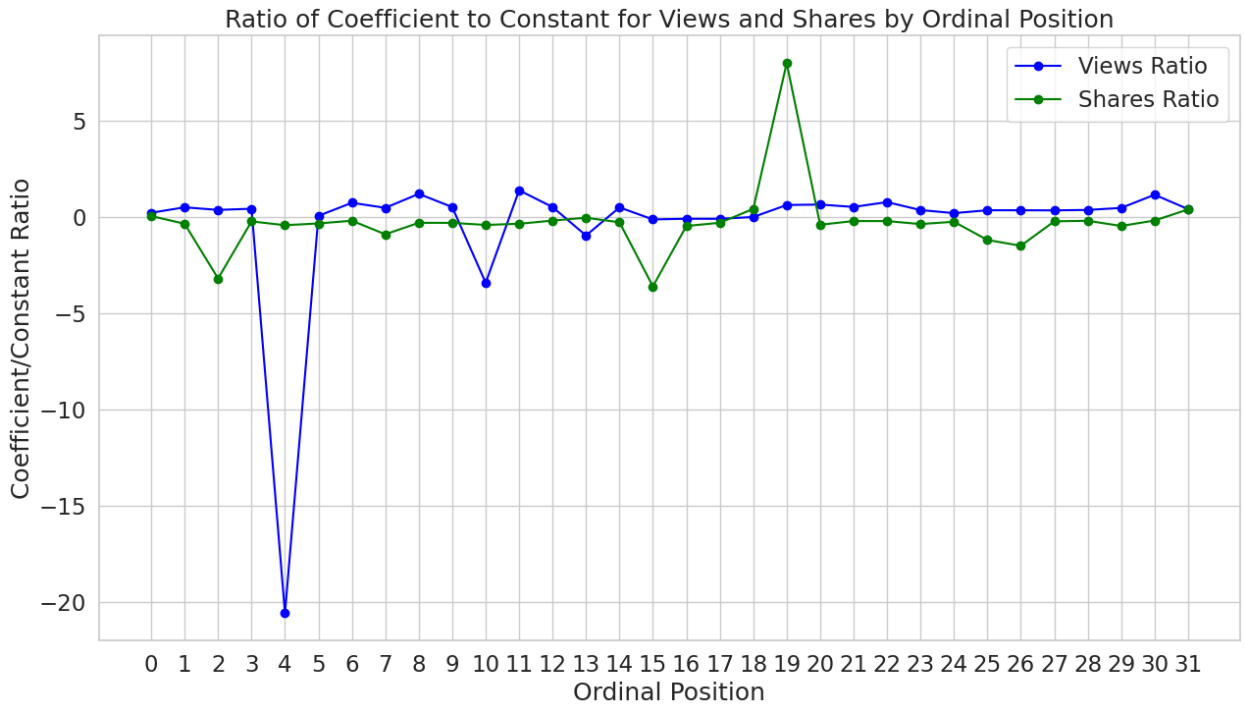
(d) Number of posts shared. F-stat: 0.271, p-value: 0.982

Notes: On average, treated users were exposed to less toxic content, but were also less active on the platform. The decrease in toxic shares is not as large as the total decrease in shares. Stickiness in sharing behavior with respect to toxic is explained by the structural model.

Figure 4: Treatment effects on post embeddings viewed and shared



(a) TE on embedding dimensions of posts viewed and shared



(b) Ratio of TE and control mean of embedding dimensions of posts viewed and shared

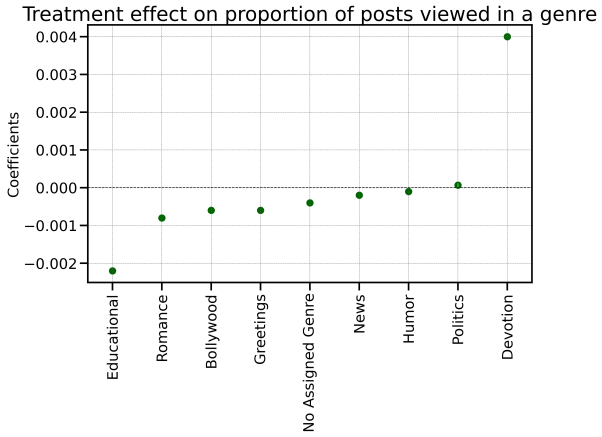
Notes: These graphs shows that on average, the treatment effect was larger for posts viewed than for posts shared. This is true even when the treatment effects are normalized according to the embedding values of posts viewed and shared by control users. This suggests that even though treatment affects content exposure, users are more likely to share content that they prefer anyway.

Table 2: Treatment effects on overall platform usage

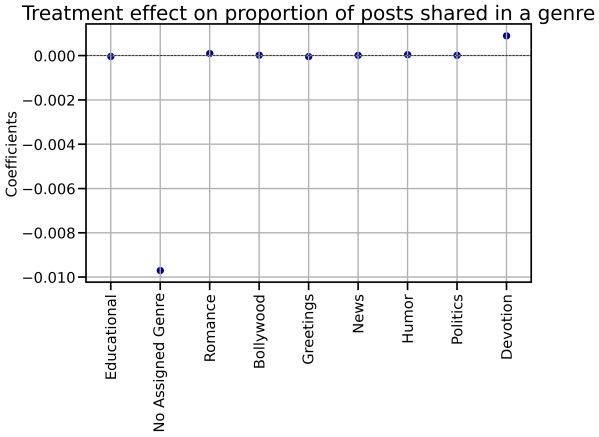
	Number of Logins	Time Spent (in hours)	Number of Posts Viewed	Time Spent per Post	Number of Posts Shared	Shares-to-Views Ratio
Treatment	-1.268**	-2.506**	-34.619**	-0.054**	-6.326**	-0.115**
	(0.042)	(0.573)	(2.170)	(0.002)	(0.203)	(0.007)
Mean DV in control group	21.534**	7.049**	243.442**	0.128**	18.267**	0.264**
	(0.021)	(0.572)	(1.337)	(0.001)	(0.129)	(0.004)
Obs	236288	236288	236288	236288	236288	236288

Notes: This table provides negative and significant estimates of effects on total number of posts viewed and shared, and total time spent on the platform. Negative treatment effect on total number of logins, suggesting differential attrition by treatment. Robust standard errors in parenthesis. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

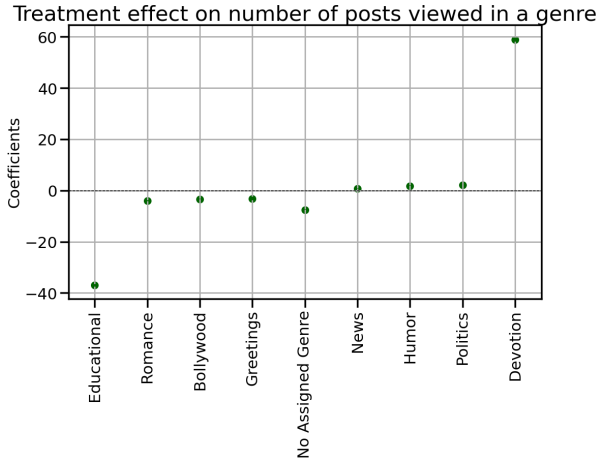
Figure 5: Treatment effects on viewing and sharing content from various genres



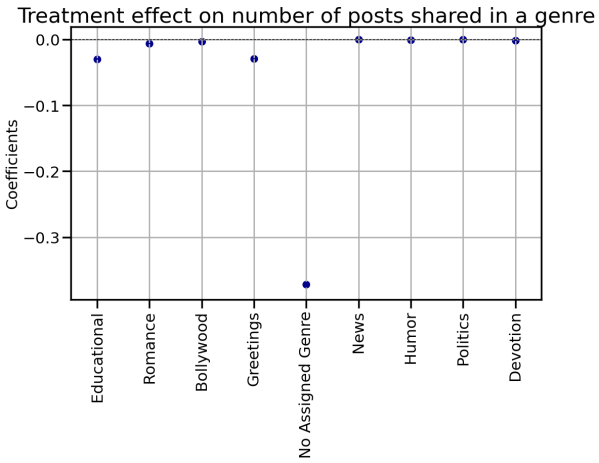
(a) Proportion posts viewed. F-stat: 6122.932, p-value: 0.000



(b) Proportion posts shared. F-stat: 1317.659, p-value: 0.000



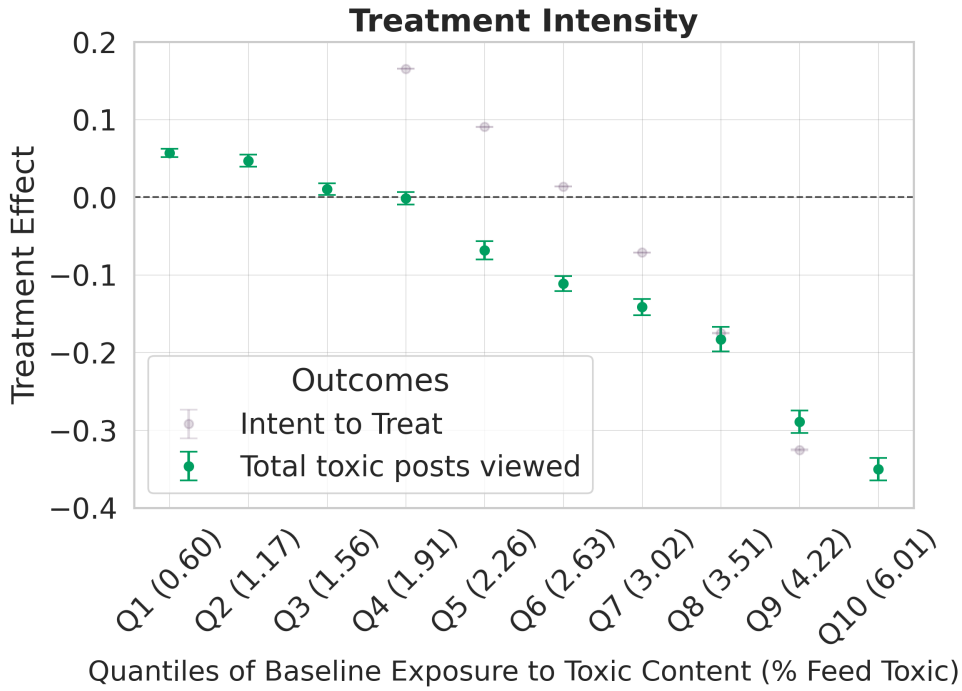
(c) Number of posts viewedF-stat: 451.112, p-value: 0.000



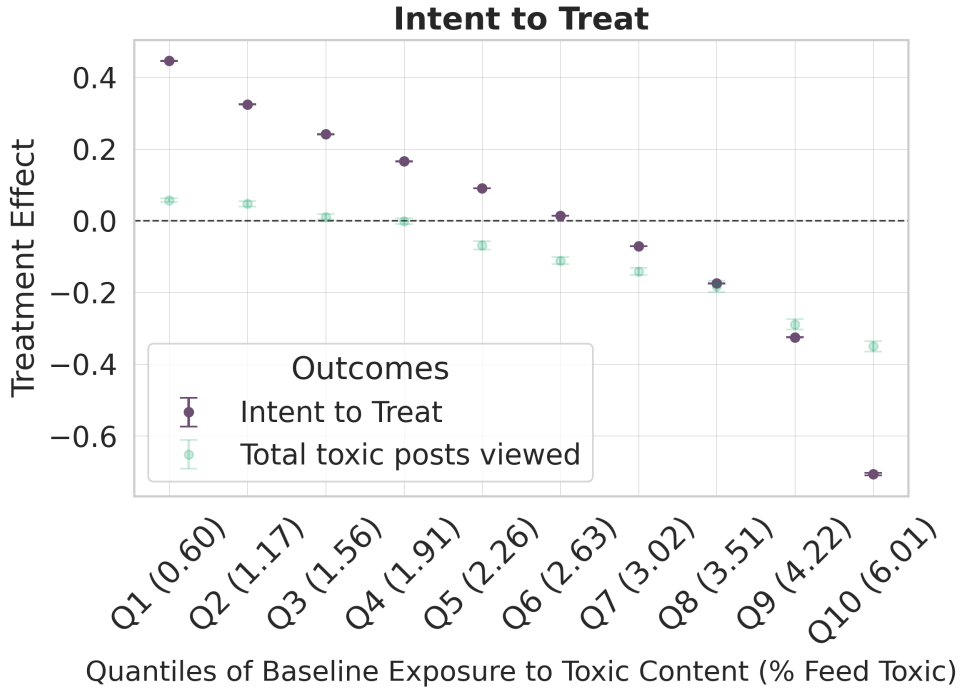
(d) Number of posts shared. F-stat: 276.637, p-value: 0.000

Notes:

Figure 6: Treatment intensity and intent-to-treat for number of toxic posts viewed



(a) Intent-to-treat is the difference between proportion of toxic posts viewed and the *predicted* proportion of toxic posts viewed



(b) Treatment intensity is the heterogenous treatment effect on number of toxic posts viewed

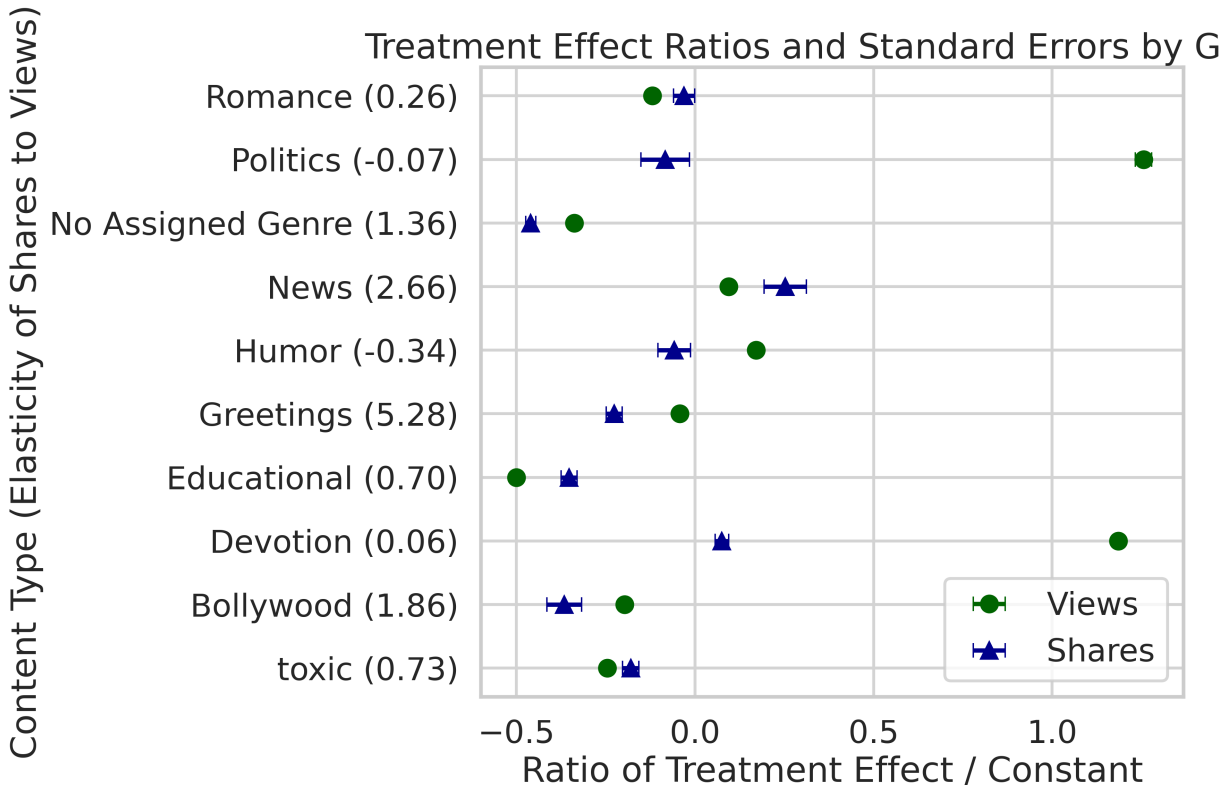
Notes: This graph shows that the treatment intensity is larger for users with lower toxic exposure at baseline. The treatment intensity is much smaller for users with highly toxic feed. The intent-to-treat, computed as the difference between proportion of toxic posts viewed and the expected proportion of toxic posts viewed, almost always has the same sign as the treatment intensity. The intent-to-treat estimates are larger than the treatment intensity estimates because the total number of posts viewed is held constant for ITT computation. Error bars are computed ar 5% level of significance.

Figure 7: Comparison of means for key outcomes across treatment and control



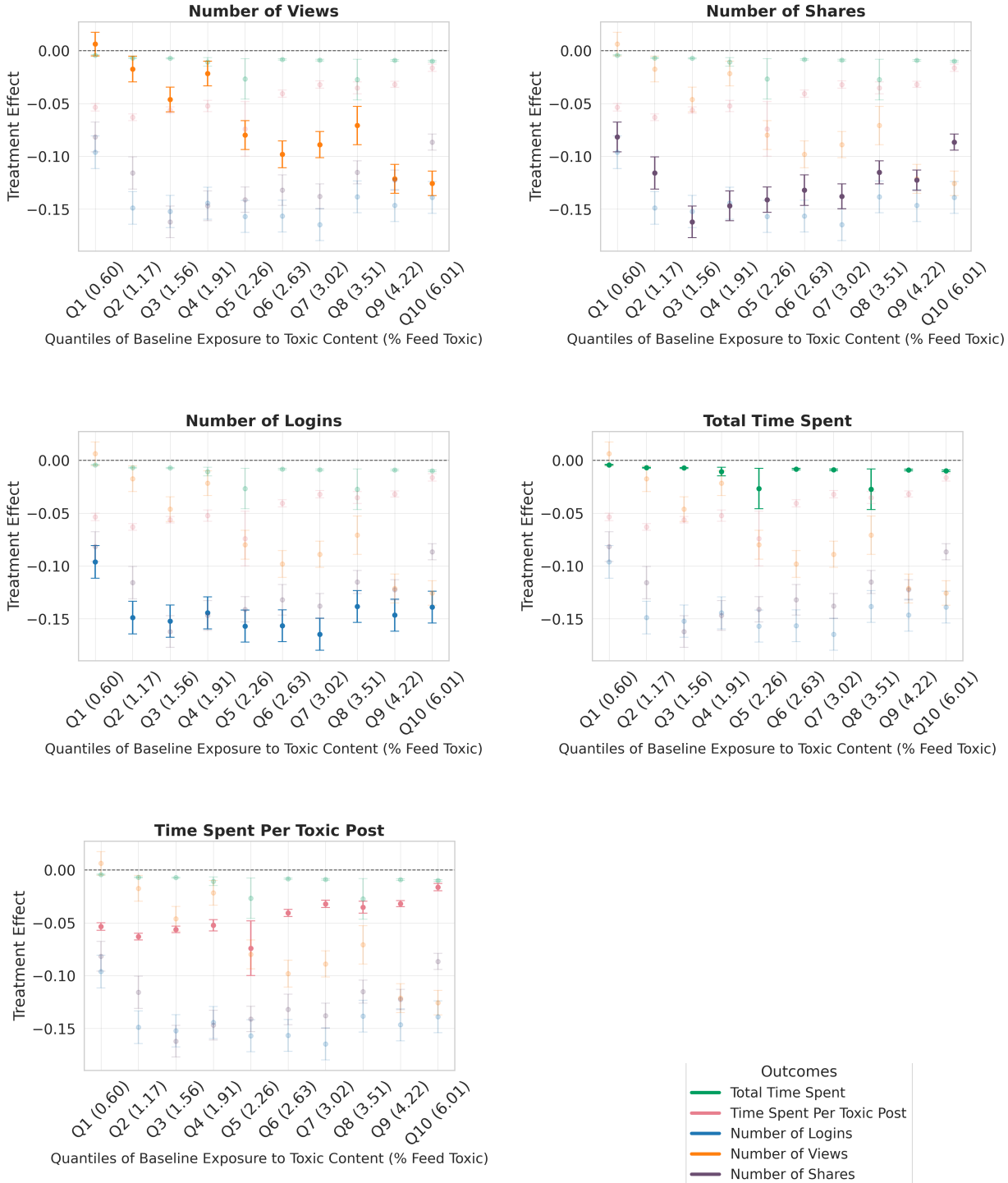
Notes: On average, treated users were exposed to less toxic content, but were also less active on the platform. The decrease in toxic shares is not as large as the total decrease in shares. Stickiness in sharing behavior with respect to toxic is explained by the structural model.

Figure 8: Elasticity of sharing different types of content



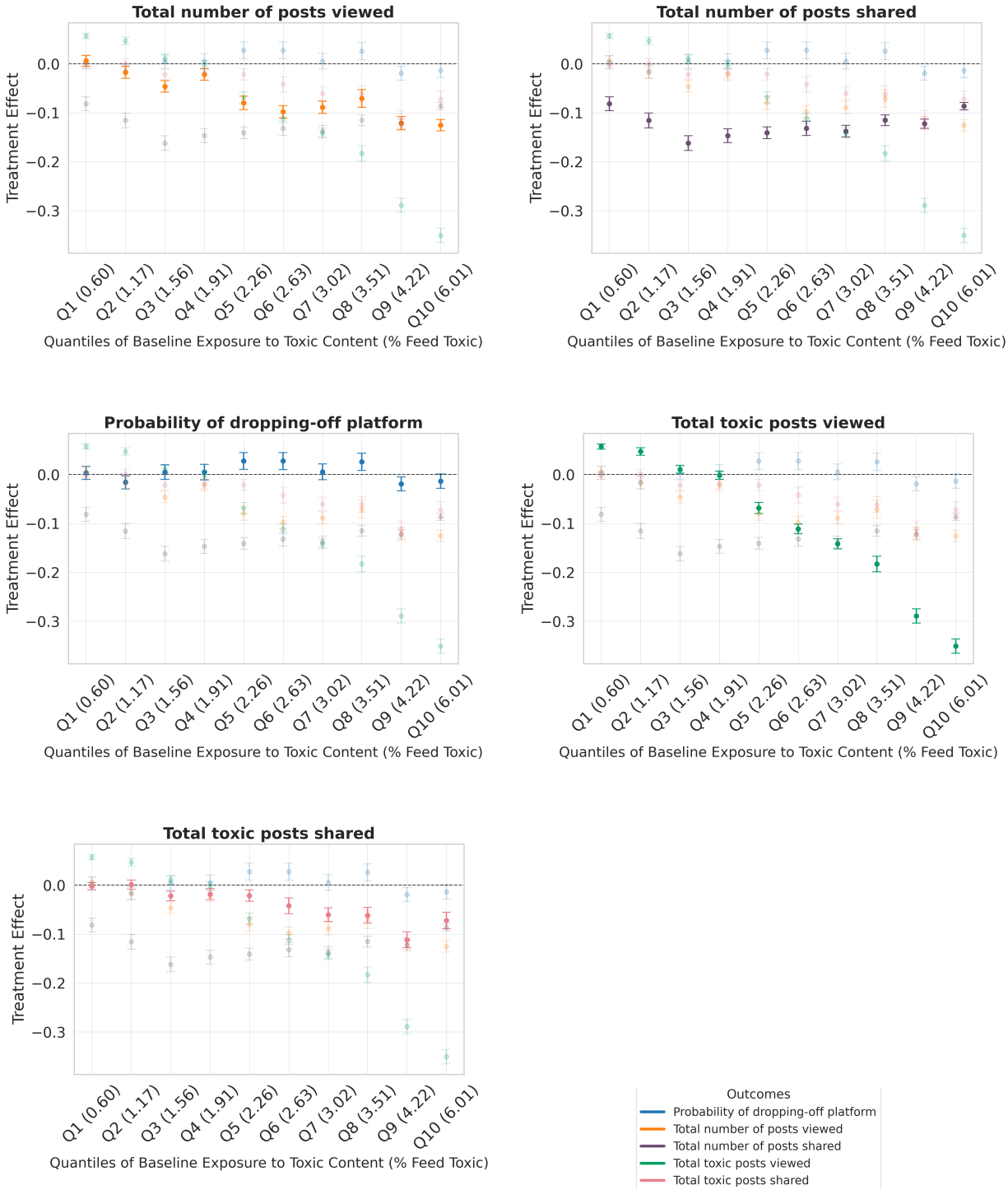
Notes: Elasticity of sharing behavior with respect to exogenous content exposure due to treatment. Elasticities are computed by dividing the purple marker with the green marker. Relevant elasticities shown in parenthesis for each content type.

Figure 9: Heterogeneous treatment effects on overall behavior



Notes: These plots show that even though the treatment effect on the number of times users log-in to the platform does not differ by heterogenous characteristics, the treatment effect on the number of posts viewed and shared is significantly lower for users with higher exposure to toxic posts at baseline. Total time spent on the platform also does not differ by toxicity of user's feed at baseline, but the treatment effect on time spent per toxic post is significantly lower (in absolute terms) for users with higher exposure to toxic posts at baseline. All variables were aggregated at user level, and standardized as z-scores.

Figure 10: Heterogeneous treatment effects on toxic behavior



Notes: These plots show that the treatment effect on the number of toxic posts viewed and shared is significantly lower for users with higher exposure to toxic posts at baseline. This is true despite the fact that the treatment effect on the number of posts viewed and shared is significantly lower for users with higher exposure to toxic posts at baseline. All variables were aggregated at user level, and standardized as z-scores.

Table 3: User characteristics correlated with the probability being exposed to more toxic content at baseline

Variable	Coefficient	Interaction Coefficient
Treatment Effect	-0.107	N/A
	(0.076)	N/A
Number of Views (Baseline)	0.0000***	-0.0000
	(0.0000)	(0.0000)
Number of Shares (Baseline)	-0.0007***	-0.0001
	(0.0001)	(0.0001)
Toxic Shares (Baseline)	0.0055***	0.0003
	(0.0003)	(0.0007)
User Gender	-0.0197**	0.0016
	(0.0074)	(0.0150)
Days since account created	-0.0002***	0.0000
	(0.0000)	(0.0000)
User Age	0.0007	0.0042*
	(0.0009)	(0.0019)
Proportion content viewed on weekends	-0.4282***	0.0200
	(0.0193)	(0.0397)
Proportion content shared during daytime	0.1780***	0.0021
	(0.0093)	(0.0183)
Share of views in Bollywood Genre	-2.0971***	-0.0943
	(0.0363)	(0.0728)
Share of views in Devotion Genre	-1.7171***	-0.0868
	(0.0333)	(0.0664)
No Assigned Genre	-1.7924***	-0.0741
	(0.0591)	(0.1198)
Share of views in Greetings Genre	-3.6599***	-0.0440
	(0.0284)	(0.0584)
Share of views in Humor Genre	6.6810***	-0.1752
	(0.0598)	(0.1151)
Share of views in News Genre	3.6608***	0.2125
	(0.0748)	(0.1476)
Share of views in Politics Genre	1.5813***	0.1854
	(0.2788)	(0.6209)
Share of views in Romance Genre	-1.0792***	-0.0192
	(0.0352)	(0.0717)

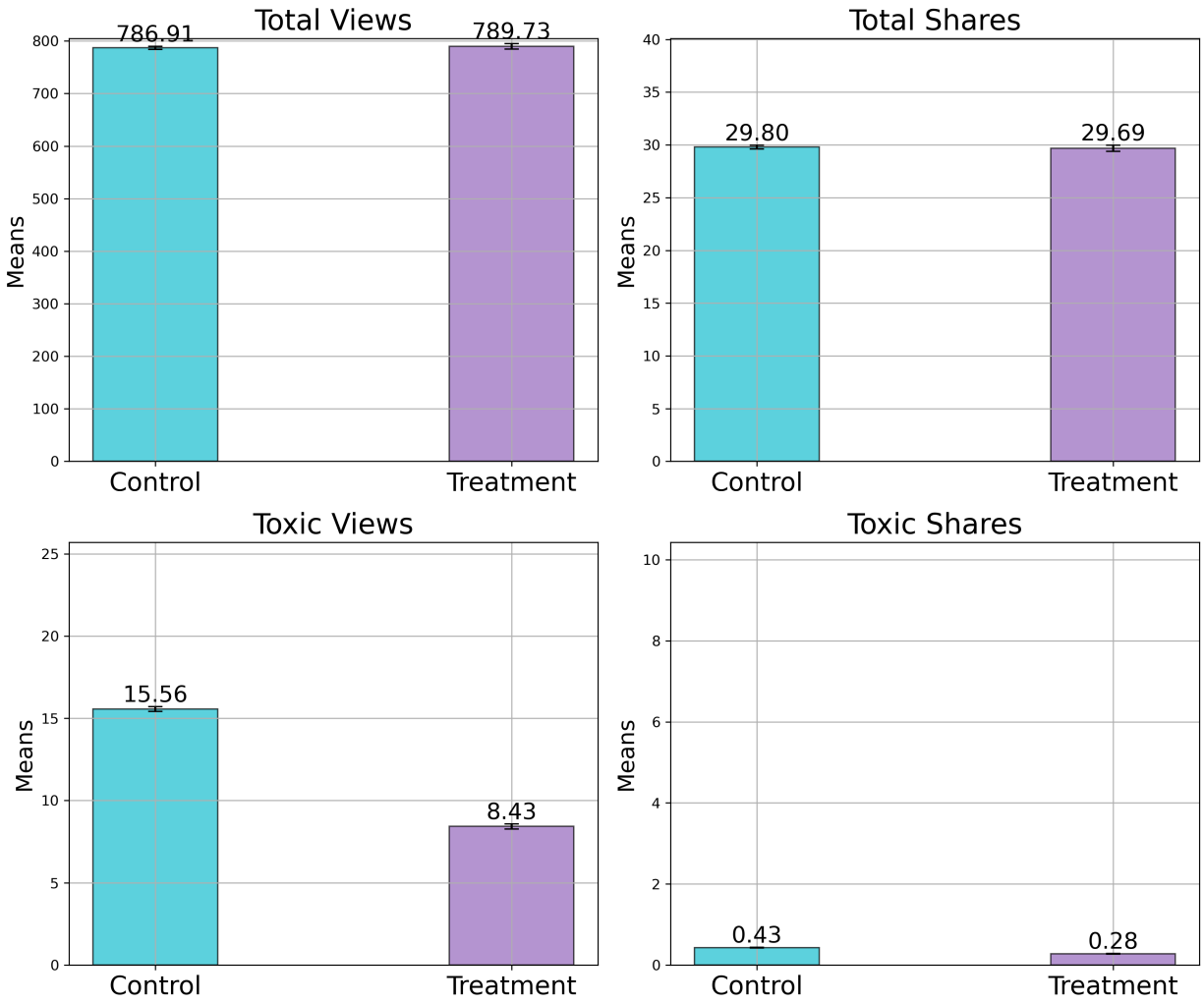
Notes: This Table shows that, conditional on observable user characteristics, treatment assignment is not correlated with the probability of leaving the platform. This also shows that the treatment does not differentially impact the probability of leaving the platform, for given observable user characteristics, in any economically significant way. This means that the treated leavers are not systematically different from the control leavers. These results are obtained by estimating the regression equation $\mathbb{1}_i(\text{left platform} = \text{yes}) = \beta_0 + \beta_1 D_i + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \sum_c \beta_{1c} D_i \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where $\mathbb{1}_i(\text{left platform} = \text{yes}$ is an indicator taking value 1, when user i leaves the platform. Column (1) reports estimated β_c's, while column (2) reports estimated β_{1c}'s. Standard errors are robust at user level. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Table 4: User characteristics correlated with the probability of leaving the platform

Variable	Coefficient	Interaction Coefficient
Treatment Effect	0.010	N/A
	(0.011)	N/A
Number of Views (Baseline)	-0.0000***	0.0000
	(0.0000)	(0.0000)
Number of Shares (Baseline)	-0.0000	0.0000
	(0.0000)	(0.0000)
Toxic Shares (Baseline)	0.0000	-0.0000
	(0.0000)	(0.0001)
Toxic Views (Baseline)	-0.0001	-0.0004
	(0.0004)	(0.0008)
User Gender	-0.0062***	-0.0026
	(0.0011)	(0.0024)
Days since account created	0.0000***	-0.0000
	(0.0000)	(0.0000)
User Age	0.0000	0.0001
	(0.0001)	(0.0003)
Proportion content viewed on weekends	-0.0004	0.0050
	(0.0029)	(0.0062)
Proportion content shared during daytime	-0.0010	0.0028
	(0.0013)	(0.0028)
Share of views in Bollywood Genre	0.0094	-0.0093
	(0.0053)	(0.0103)
Share of views in Devotion Genre	-0.0107*	-0.0001
	(0.0045)	(0.0092)
No Assigned Genre	0.0088	0.0281
	(0.0094)	(0.0200)
Share of views in Greetings Genre	0.0044	-0.0131
	(0.0046)	(0.0093)
Share of views in Humor Genre	0.0206**	0.0085
	(0.0079)	(0.0174)
Share of views in News Genre	0.0067	-0.0178
	(0.0066)	(0.0137)
Share of views in Politics Genre	0.0150	-0.0603
	(0.0405)	(0.0651)
Share of views in Romance Genre	0.0172**	0.0066
	(0.0055)	(0.0112)

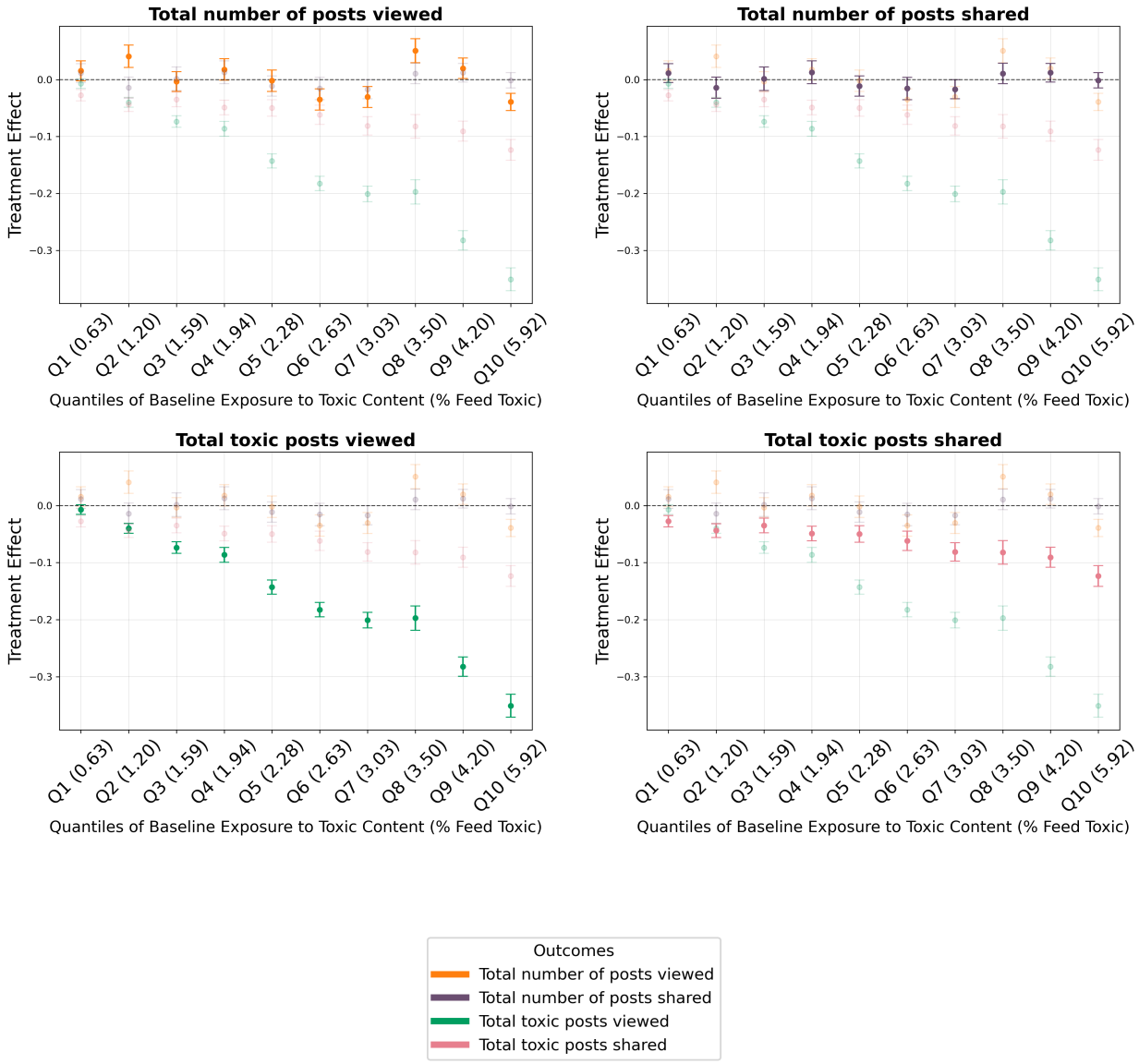
Notes: This Table shows that, conditional on observable user characteristics, treatment assignment is not correlated with the probability of leaving the platform. This also shows that the treatment does not differentially impact the probability of leaving the platform, for given observable user characteristics, in any economically significant way. This means that the treated leavers are not systematically different from the control leavers. These results are obtained by estimating the regression equation $\mathbb{1}_i(\text{left platform} = \text{yes}) = \beta_0 + \beta_1 D_i + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \sum_c \beta_{1c} D_i \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where $\mathbb{1}_i(\text{left platform} = \text{yes}$ is an indicator taking value 1, when user i leaves the platform. Column (1) reports estimated β_c's, while column (2) reports estimated β_{1c}'s. Standard errors are robust at user level. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Figure 11: Comparison of means for key outcomes across treatment and control stayers



Notes: This graph shows that on average, there are no significant differences in platform engagement (total views and shares) between *stayers* in treated and control groups. On average, treated stayers were exposed to less toxic content, but were also less active on the platform. The percentage decrease in toxic shares is an order of magnitude larger than the total decrease in shares. These findings imply that stayers may face increased returns from sharing toxic content, when the treatment randomly reduces the number of toxic posts they are exposed to. These findings, combined with the fact that the treated leavers do not differ from control leavers along observable characteristics, suggests that the treatment effects are not driven by selection.

Figure 12: Heterogeneous treatment effects on toxic behavior among stayers



Notes: These plots show that the treatment effect on the number of toxic posts viewed and shared is significantly lower for stayers with higher exposure to toxic posts at baseline. This is true despite the fact that the treatment effect on the number of posts viewed is significantly lower for stayers with highest exposure to toxic posts at baseline. All variables were aggregated at user level, and standardized as z-scores.

Table 5: 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 DV 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. The coefficient estimates are obtained from stacking regressions of (non-)toxic shares on (non-)toxic views. The statististical test of equality of coefficients could not reject the hypothesis that the coefficients from the two regressions are equal. Robust standard errors in parenthesis. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$

Table 6: Testing identifying assumption in structural model

	(1)	(2)	(3)
	Probability of sharing toxic post during intervention period		
Probability of sharing toxic post at baseline	0.209***		1.105***
	(0.022)		(0.119)
Proportion of toxic posts among first half of posts shared at baseline		0.478***	
		(0.055)	
Mean DV in control group	4.52e-08	8.22e-09	3.48e-08
	(0.030)	(0.034)	(0.032)
N	28384	28384	28384

Notes: This Table tests the identifying assumption $s^t_{i,0} = s^t_{i,1}$ in the control group. Column (3) shows that the measurement error corrected estimates of the intercept and slope coefficient are very close to 0 and 1 respectively. I cannot reject the hypothesis that the measurement error corrected slope coefficient is equal to 1. All variables are standardized to have zero means. Robust standard errors in parenthesis. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$

Table 7: Structural estimation with measurement error correction

	(1)	(2)	(3)
	Difference in toxic views at baseline (half = 2)	Difference in toxic shares (current - baseline)	
Difference in toxic views at baseline (half = 1)	0.558***	-0.142*	
	(0.00444)	(0.0607)	
Difference in toxic views at baseline			-0.155*
			(0.0704)
Constant	-42.27***	-13.96*	-15.20*
	(0.428)	(5.831)	(6.765)
N	63041	63041	63041

Notes: This Table provides estimates for the structural parameter θ in the model of sharing behavior. θ is the parameter that captures 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. The model then shows that the rate at which users update behavior according to the perceived behavior of others is also θ and is very low. Column (1) shows results of an OLS regression of differences between probability of viewing toxic and non-toxic content, computed using only the second half of posts viewed by user at baseline, when they were arranged in a random order ($half2$). The independent variable in Column (1) is difference between proportion of toxic and non-toxic posts viewed at baseline, computed using only the first half of posts viewed by user at baseline, when they were arranged in a random order ($half1$). The first stage relationship in Column (1) is positive, significant, but is not close to 1. Column (2) shows results of a 2SLS regression of difference between baseline and intervention period in differences between probability of sharing toxic and non-toxic content. In Column (2), the independent variable is $half1$, which was instrumented with $half2$. Column (3) estimates the model with measurement error correction, where the correlation between $half1$ and $half2$ serves as the reliability measure for the main right hand side variable. Estimated slope coefficient estimates γ_1 is always negative and statistically significant. Estimated θ is therefore, positive. Baseline period is December, 2022 and intervention period data spans from February 10, 2023 and March 10, 2023. Robust standard errors in parenthesis. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

A Supplementary Figures

Figure A.1: 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. Users can share, comment, like, or download the post to their phone. Sharing refers to sharing on WhatsApp, and not on the platform. A user can see other users who liked and commented on a post, but not the users who shared the post.

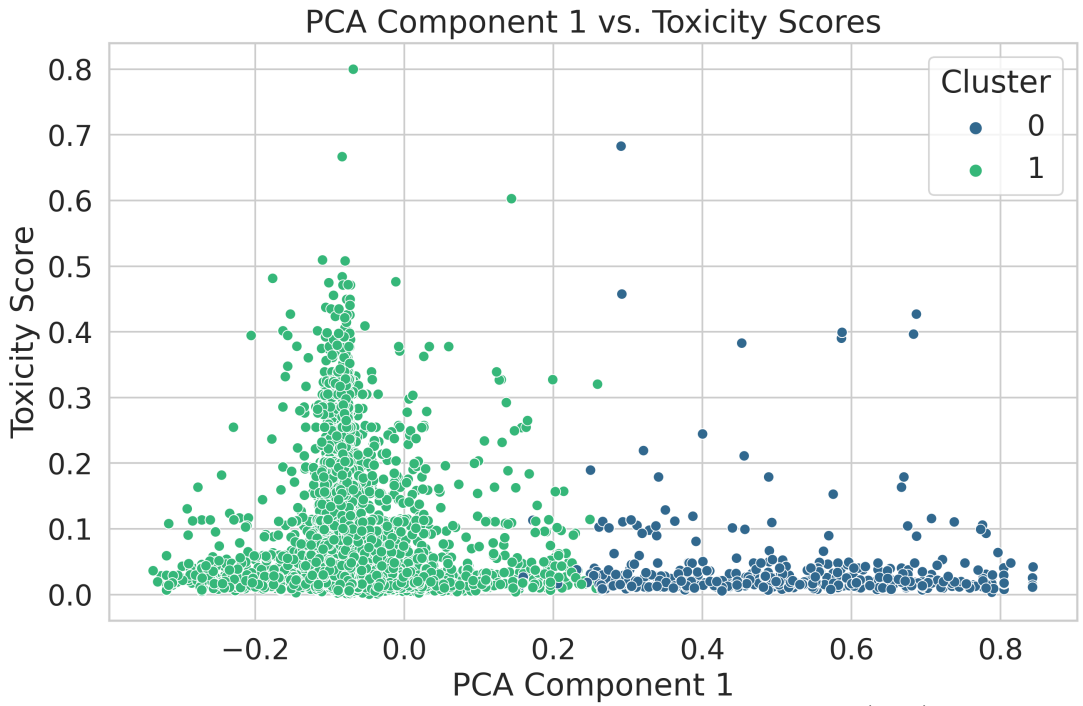
Figure A.2: Word clouds depicting words associated with highly toxic posts



(a) High Toxicity (b) Low Toxicity

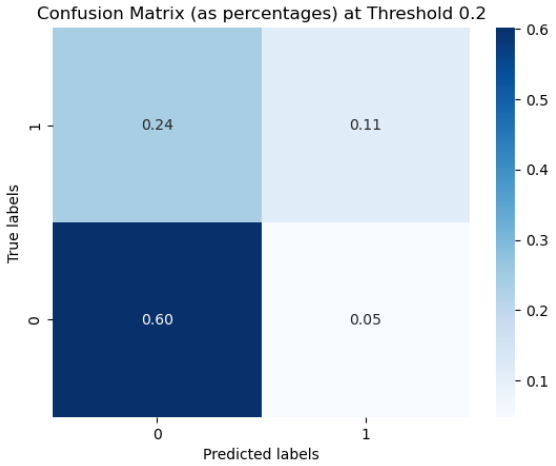
Notes: This Figure shows word clouds constructed using TF-IDF vectorizer on posts classified into high and low toxicity categories respectively. Cut-off to classify posts into high and low toxicity categories is 0.2. The figure demonstrates overlap in words pertaining to religion in both categories, 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 A.3: Validating toxicity score threshold to construct binary outcome

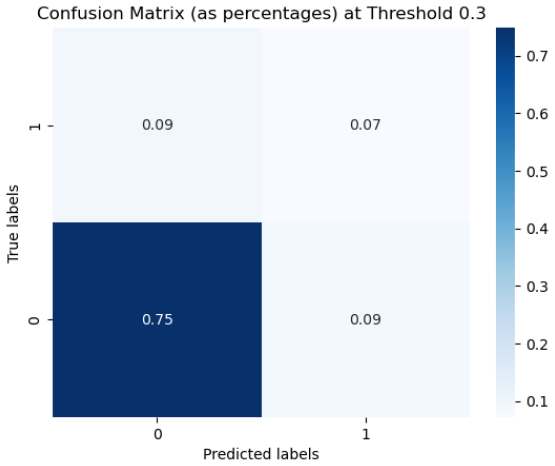


Notes: X-axis plots the first principal component of words in each document (post). Y-axis corresponds to the toxicity score for the corresponding documents (posts). Toxicity scores for most posts in the first (blue) cluster of a PCA representation of words rarely exceed 0.2. 0.2 is chosen as the appropriate threshold for the binary outcome variable, indicating whether a post is toxic or not.

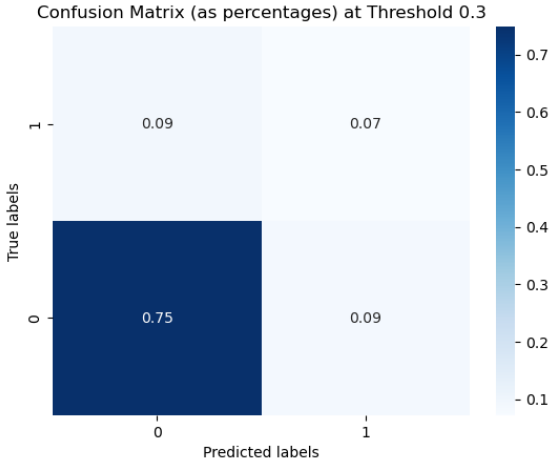
Figure A.4: Confusion matrices for different cut-offs in toxicity scores



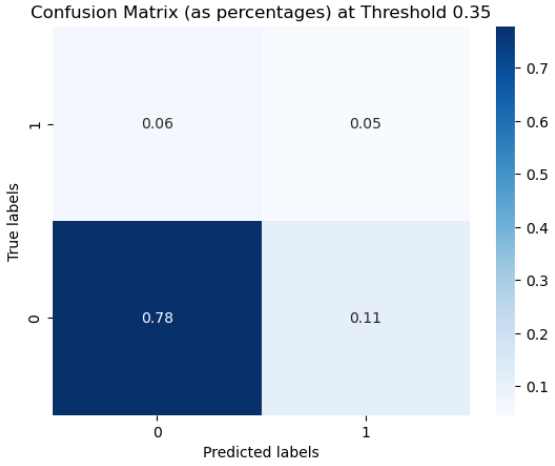
(a) Toxicity score cut-off = 0.2



(b) Toxicity score cut-off = 0.25



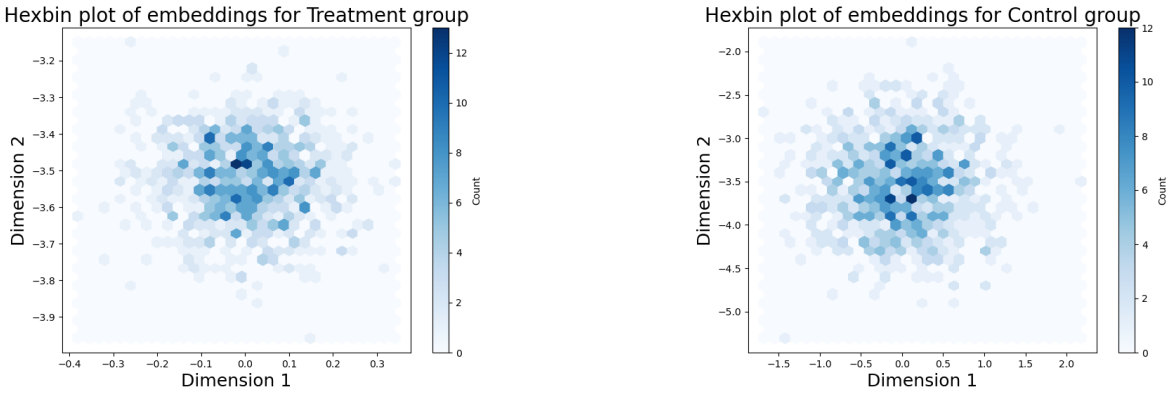
(c) Toxicity score cut-off = 0.3



(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: 0.2, 0.25, 0.3, 0.35. User posts were granted 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 human annotators hired at Brown University.

Figure A.5: Joint distribution of dimensions in simulated embeddings for treatment and control groups

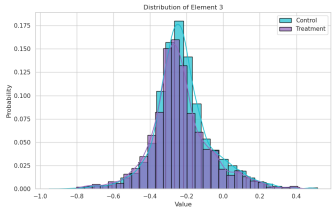


(a) Treatment Group

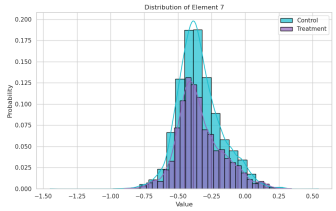
(b) Control Group

Notes: This graph shows that the two dimensions (components) of the simulated embedding vector can be expected to be more concentrated around the mean for the treated group. This is expected by application of the central limit theorem as the treatment embeddings were drawn each day from a given sample of control embeddings during the intervention period. The algorithm used to simulate the embeddings for both treatment and control groups uses Singular Value Decomposition to factorize the matrix of engagement in the control group. Embeddings from the treatment group were uniformly drawn from an epsilon ball centred around the mean control embedding. Details of the simulation are contained in Appendix E.

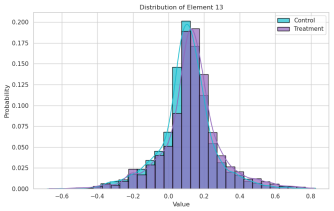
Figure A.6: Treatment effect on embeddings across various Dimensions



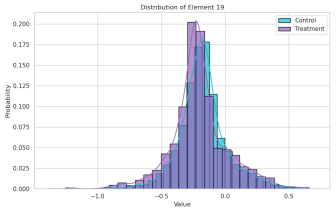
(a) 3



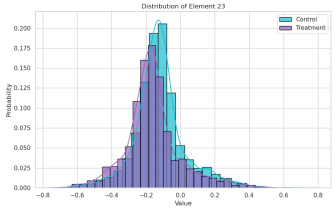
(b) 7



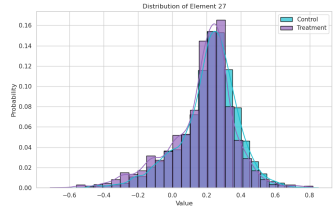
(c) 13



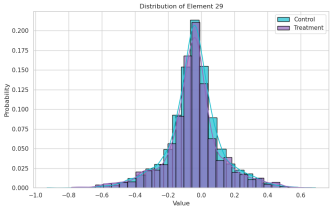
(d) 19



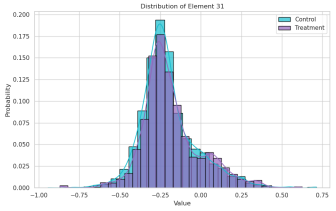
(e) 23



(f) 27



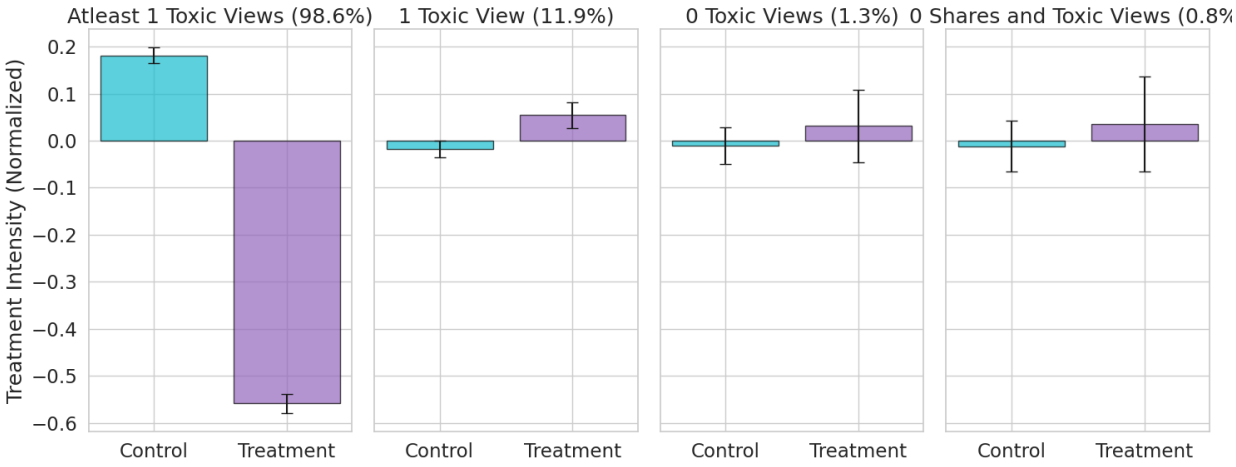
(g) 29



(h) 31

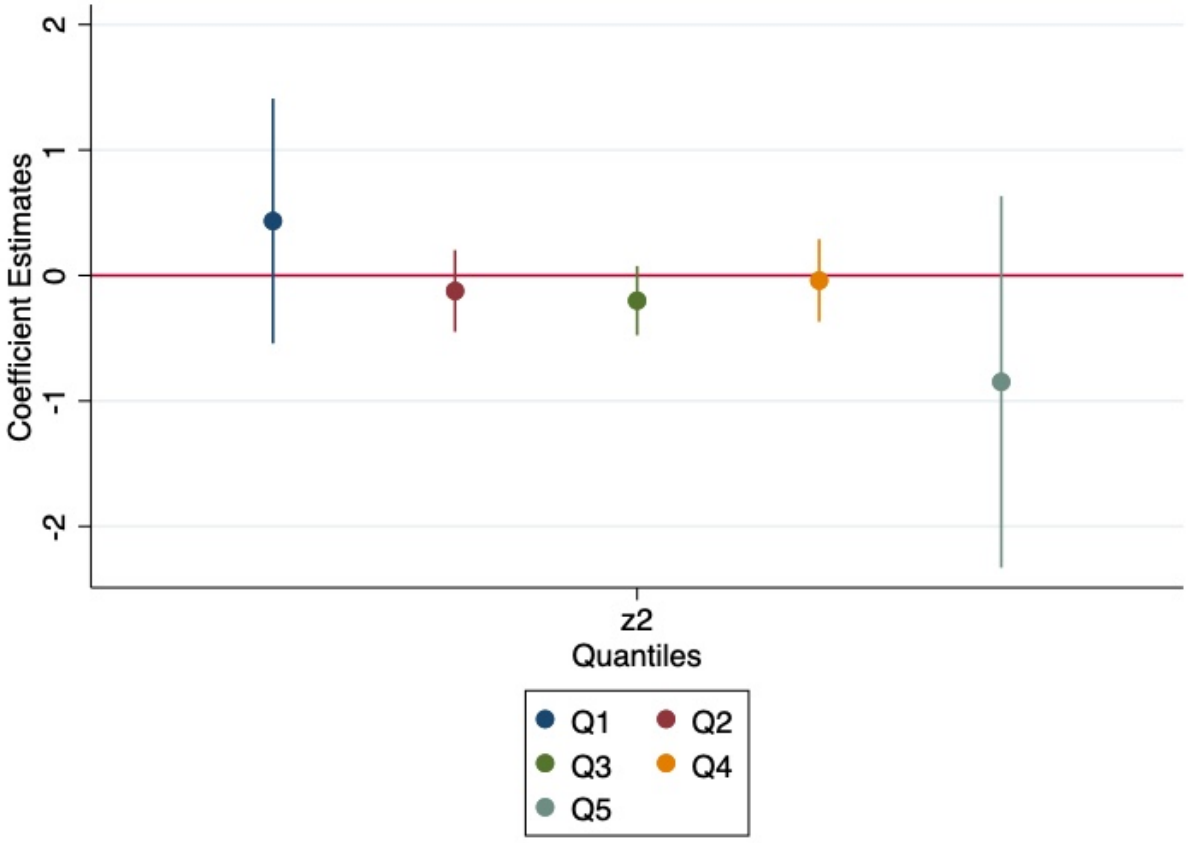
Notes: This figure shows the treatment effect on randomly selected dimension of user embeddings for the treatment and control groups. On average, the user embeddings for the treatment group are more concentrated around the mean than the control group, as predicted by the simulated embeddings. Embedding dimensions were rarely correlated with human intelligible topics or content genres.

Figure A.7: Treatment intensity for different types of users

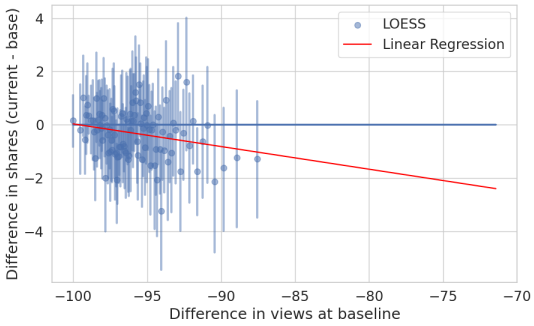


Notes: This graph shows that treatment intensity is negative for users who saw more than 1 toxic posts at baseline. User types are defined using toxic exposure at baseline, and the percentage users of a particular type is shown in paranthesis. Treatment intensity is computed as the difference between actual exposure to toxic content and predicted exposure to toxic content using baseline characteristics. The distribution of the constructed variable is normalized around the mean over both treated and control users.

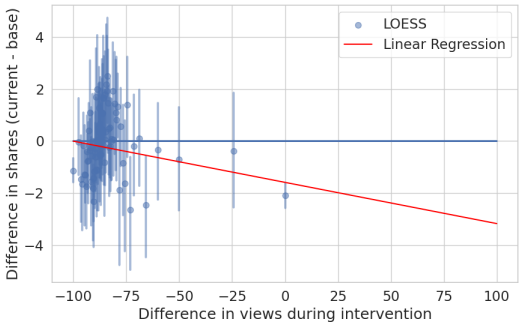
Figure A.8: Testing simplifying assumotion in action-signalling model



Notes: This graph fails to reject the hypothesis that heterogenous users update their behavior at different 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.

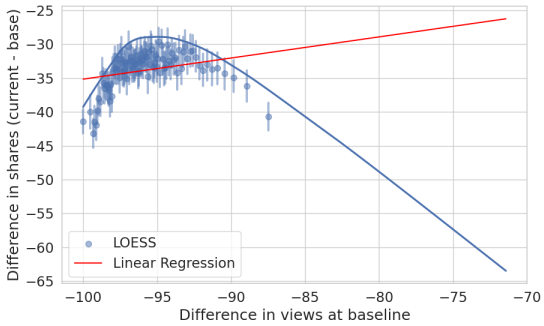


(a) Baseline views and intervention period shares in the treatment group

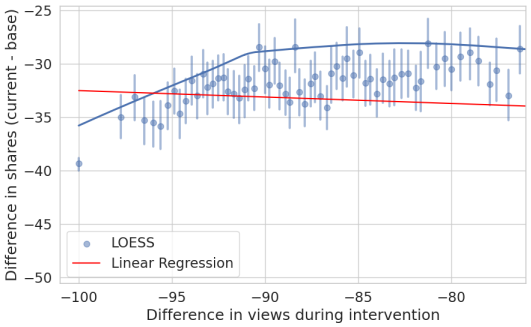


(b) Intervention period views and shares in the treatment group

Figure A.9: Exposure to toxic content and engagement with toxic content in and baseline and in the same time period among *control* users



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



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

B Supplementary Tables

Table B.1: Balance in observable characteristics among stayers

Variable	Coefficient	Standard Error
state: gujarat	-0.016	0.022
city: auraiya	-0.059	0.057
city: bhiwani	-0.011	0.053
city: durg	0.007	0.045
city: gwalior	-0.058	0.049
city: jalandhar	-0.019	0.052
city: kutch	0.051	0.057
city: mahendragarh	0.003	0.061
city: mandsaur	-0.025	0.049
city: rajnandgaon	0.008	0.055
city: ranga reddy	0.025	0.049
city: ratlam	-0.004	0.051
city: sagar	-0.039	0.050
city: samastipur	-0.084	0.053
city: surat	0.011	0.041
city: udaipur	0.040	0.053
city: udham singh nagar	0.035	0.053
Male	0.005	0.004
18-24	-0.003	0.006
45+	-0.015	0.015

Notes: The table displays a randomly chosen subset of characteristics and corresponding coefficients from one regression with all the characteristics on the right hand side, and treatment status on the left hand side. None of the coefficients are individually or jointly significant when this regression was estimated using a sample of stayers

Table B.2: User characteristics correlated with the probability of leaving the platform conditional on users being extreme at baseline

Variable	Coefficient	Interaction Coefficient
Treatment Effect	0.032	N/A
	(0.032)	N/A
Number of Views (Baseline)	-0.0000***	0.0000
	(0.0000)	(0.0000)
Number of Shares (Baseline)	-0.0000	0.0000
	(0.0000)	(0.0000)
Toxic Shares (Baseline)	0.0002	-0.0001
	(0.0002)	(0.0003)
User Gender	-0.0158***	-0.0099
	(0.0035)	(0.0074)
Days since account created	0.0000***	-0.0000
	(0.0000)	(0.0000)
User Age	-0.0002	-0.0000
	(0.0003)	(0.0007)
Proportion content viewed on weekends	-0.0133	0.0202
	(0.0085)	(0.0187)
Proportion content shared during daytime	0.0025	0.0049
	(0.0041)	(0.0090)
Share of views in Bollywood Genre	-0.0617***	-0.0169
	(0.0137)	(0.0278)
Share of views in Devotion Genre	-0.0821***	0.0024
	(0.0131)	(0.0266)
No Assigned Genre	-0.0308	0.0398
	(0.0261)	(0.0516)
Share of views in Greetings Genre	-0.1031***	-0.0521*
	(0.0126)	(0.0252)
Share of views in Humor Genre	0.2879***	0.1203
	(0.0408)	(0.1051)
Share of views in News Genre	0.1262***	-0.0591
	(0.0278)	(0.0585)
Share of views in Politics Genre	0.0640	-0.1950
	(0.1455)	(0.2219)
Share of views in Romance Genre	0.0042	0.0246
	(0.0169)	(0.0341)

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, in any economically significant way. This means that the treated leavers are not systematically different from the control leavers. These results are obtained by estimating the regression equation $\mathbb{1}_i(\text{left platform} = \text{yes}) = \beta_0 + \beta_1 D_i + \sum_c \beta_c \mathbb{1}_i(\text{user characteristic} = c) + \sum_c \beta_{1c} D_i \mathbb{1}_i(\text{user characteristic} = c) + \varepsilon_i$, where $\mathbb{1}_i(\text{left platform} = \text{yes}$ is an indicator taking value 1, when user i leaves the platform. Column (1) reports estimated β_c's, while column (2) reports estimated β_{1c}'s. Standard errors are robust at user level. $p < .0001^{***}$, $p < .01^{**}$, $p < .05^{*}$.

Table B.3: Structural estimates using OLS regression across heterogenous users

	(Q1)	(Q2)	(Q3)	(4)	(Q5)
	Difference in Toxic Shares (Current - Baseline)				
Difference in Toxic Views at	0.434	-0.124	-0.201	-0.0402	-0.848
Baseline using one half of posts	(0.498)	(0.166)	(0.141)	(0.168)	(0.756)
Constant	42.76	-12.72	-19.59	-4.697	-78.27
	(49.08)	(16.15)	(13.50)	(15.87)	(68.95)
N	12609	12623	12596	12613	12600

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 for treated users of different types, based on their exposure to toxic content at baseline. Robust standard errors in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table B.4: Structural estimates using OLS regressions

	Difference in shares (Current - Baseline)
Difference in views at Baseline	-0.085*
	(0.039)
Constant	-8.486*
	(3.745)
Obs	63270

Robust standard errors in paranthesis.

Notes: Dependent variable is differences in differences between probability of sharing toxic and non-toxic content, between intervention period and baseline, for treated users. The explanatory variables are constructed by averaging differences between proportion of toxic and non-toxic posts viewed by treated users. θ estimated in the treated sample suffers from attenuation bias due to measurement error. Robust standard errors in parentheses. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

Table B.5: Structural estimates using OLS regression and control sample

	Difference in shares (Current - Baseline)	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)
Obs	193437	193437

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. $p < 0.05^{*}, p < 0.01^{**}, p < 0.001^{***}$.

C Predictions from General Model

I compute the Subgame Perfect Equilibria using backward induction, first solving for users' optimal choice of sharing probabilities, when the user has viewed a toxic or non-toxic post. At this point in the game, the user knows the total number of posts she will view. I also assume that the user knows the distribution of toxic posts that the algorithm assigns her, q^t User optimally chooses (s^t, s^n), to maximize the utility function in (14). For some quadratic functions $u(\cdot)$ and $c(\cdot)$, user's first order condition implies

$$\frac{\partial u(s^t; q^t, \beta^t, \alpha)}{\partial s^t} = \frac{\partial c(s^t - \theta \bar{s}^t - (1-\theta)p^t)}{\partial s^t} \tag{29}$$

$$\xi_1(\alpha) + \xi_2(\alpha) \cdot s^t = \kappa_1(c) \cdot (s^t - \theta q^t - (1-\theta)p^t) \tag{30}$$

because, $u(\cdot)$ is assumed to be weakly concave and quadratic in s^t, it's first derivative is a linear combination of s^t with some unknown constants, $\xi_1(\alpha)$ with $\xi_2(\alpha) \leq 0$. These constants necessarily depend on α, which denotes the user's consumption value from sharing a post. Similarly, the constant $\kappa_1(c) > 0$ is obtained from the convexity assumption on quadratic function $c(\cdot)$. Then, user's best response functions are characterized from the first order conditions as

$$s^t = \frac{\xi_1(\alpha)}{\xi_2(\alpha) - \kappa_1(c)} - \theta \frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)} q^t - (1-\theta)\frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)} p^t \tag{31}$$

and similarly,

$$s^n = \frac{\xi_1(\alpha)}{\xi_2(\alpha) - \kappa_1(c)} - \theta \frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)} q^n - (1-\theta)\frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)} p^n \tag{32}$$

$s^t, s^n \in [0, 1]$. The user's best response strategy shows that sharing behavior in equilibrium is determined by a convex combination of user's own taste for content and for sharing, as well as some reference level of toxic posts that a user is exposed to, q^t. These assumptions enable me to present some testable hypothesis that help characterizing the reduced form estimates, and the heterogeneity therein.

Proposition 9. *For $q^t \in [0, 1]$, $\xi_2(\alpha) \leq 0$, and $\kappa_1(c) \geq 0$,*

$$\frac{\partial S^t}{\partial q^t} \leq 0 \tag{33}$$

That is, the probability of sharing toxic content is increasing in the probability of being assigned such content.

Proof. For some fixed ν,

$$\frac{\partial S^t}{\partial q^t} = -\theta\nu \frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)}$$

Then, under the stated assumptions, $\xi_2(\alpha) \leq 0$, and $\kappa_1(c) \geq 0$,

$$-\theta\nu \frac{\kappa_1(c)}{\xi_2(\alpha) - \kappa_1(c)} \geq 0$$

□

The reduced form estimates show that the treatment reduces the total number of toxic posts viewed for users engaging with more toxic content at baseline. This proposition implies that for such users, the treatment is expected to reduce the total number of toxic posts shared as well. This is a direct implication of users' innate tastes for sharing toxic content, as well as the action-signalling value.

Proposition 10. *For $q^t \in [0,1]$, $\xi_2(\alpha) \leq 0$, and $\kappa_1(c) \geq 0$,*

$$\frac{\partial^2 s^t}{\partial q^t \partial \xi_2} \geq 0 \tag{34}$$

That is, the treatment effect on toxic shares is larger for users with higher marginal disutility for such content.

Proof. This is the derivative in Proposition 9 with respect to $\xi_2(\alpha)$. □

Heterogeneous treatment effects are larger for users who had lower marginal utility from sharing toxic content. That is, users with higher proclivity to toxic content, or positive taste from sharing toxic content, had smaller treatment effects. These treatment effects, however, do not disentangle the effects of user tastes from their signalling value. I estimate the effects of users innate tastes using a simpler structural model in Section **??**.

Users are assumed to know the probability distribution of being assigned toxic or non-toxic posts at the beginning of each session, even though she cannot know which posts she will view before the session starts. This information is however, sufficient to compute the total number of posts she should view, ν, to maximize her utility. Substituting optimal sharing function in the utility function to optimize,

$$\begin{aligned}
\max_{\nu} U = \nu[&u(\frac{\xi_1}{\xi_2 - \kappa_1} - \theta\frac{\kappa_1}{\xi_2 - \kappa_1}q^t - (1-\theta)\frac{\kappa_1}{\xi_2 - \kappa_1}p^t; q^t, \beta^t) \\
&+ u(\frac{\xi_1}{\xi_2 - \kappa_1} - \theta\frac{\kappa_1}{\xi_2 - \kappa_1}q^n - (1-\theta)\frac{\kappa_1}{\xi_2 - \kappa_1}p^n; q^n, \beta^n) \\
&- c(\frac{\xi_1}{\xi_2 - \kappa_1} - \theta\frac{\xi_2}{\xi_2 - \kappa_1}q^t - (1-\theta)\frac{\xi_2}{\xi_2 - \kappa_1}p^t) \\
&- c(\frac{\xi_1}{\xi_2 - \kappa_1} - \theta\frac{\xi_2}{\xi_2 - \kappa_1}q^n - (1-\theta)\frac{\xi_2}{\xi_2 - \kappa_1}p^n) \\
&- \eta^2\nu]
\end{aligned}$$

The demand for social media consumption can be measured using the number of posts (both toxic and non-toxic) users choose to view. Users can respond to the treatment by changing the total time they spend on the platform looking at posts. This implies that an experiment that assigns toxic posts randomly is costly for the firm.

Proposition 11. *When $s^t - (1-\theta)p^t \leq \theta q^t$, an intervention that randomly reduces q^t is costless to the firm.*

Proof. In appendix. □

If users' innate taste for viewing and sharing toxic content is weaker than the perceived societal tastes, an intervention that reduces q^t does not reduce overall usage of the platform. This also suggests that if this condition is not true, that is if the probability of assigning toxic posts is lower than user's innate tastes for such content, users are likely to view less toxic content

Proposition 12.

$$\frac{\partial^2 \nu}{(\partial q^t)^2} \leq 0 \tag{35}$$

That is, the demand curve for social media consumption is an inverted-U shaped curve.

Proof. This flows from the proof of Proposition 11. □

The platform or the algorithm optimally chooses (q^t, q^n) for each user by a priori maximizing ν. I measure the quantities of the attention commodity in terms of total number of posts viewed. $\nu(q^t, p^t, p^n)$ is the correct objective function for the platform to maximize, because a user who views a higher number of total posts also consumes a greater number of advertisements. This objective function, therefore, allows the firm to maximize ad-revenues.

$$\begin{aligned}\max_{q^t,q^n} \nu = \frac{1}{2\eta}\Bigg[& u(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\kappa_1}{\xi_2-\kappa_1}q^t - (1-\theta)\frac{\kappa_1}{\xi_2-\kappa_1}p^t; q^t, \beta^t) \\ & + u(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\kappa_1}{\xi_2-\kappa_1}q^n - (1-\theta)\frac{\kappa_1}{\xi_2-\kappa_1}p^n; q^n, \beta^n) \\ & - c(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\xi_2}{\xi_2-\kappa_1}q^t - (1-\theta)\frac{\xi_2}{\xi_2-\kappa_1}p^t) \\ & - c(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\xi_2}{\xi_2-\kappa_1}q^n - (1-\theta)\frac{\xi_2}{\xi_2-\kappa_1}p^n)\Bigg]\end{aligned}$$

First, consider the case of no updating, $\theta = 0$. That is, users do not update their perceptions of society based on their feeds or the toxic content they are exposed to. In this case, I show that the assignment probabilities are entirely determined by user's innate tastes for toxic content, and is not affected by the signalling value a user may derive from sharing. This is because the user continues to behave according to some predetermined preferences, in each time period.

Proposition 13. *When $\theta = 0$, and for $\zeta_2^t(\beta^t), \zeta_2^n(\beta^n) \leq 0$*

$$q^t = \frac{\zeta_1^n + \zeta_2^n - \zeta_1^t}{\zeta_2^t + \zeta_2^n} \tag{36}$$

That is, the probability of assigning toxic posts is directly proportional to users' distaste for non-toxic content, when there is no updating.

Proof. In appendix. □

In the case of no updating, users do not receive any benefit from signalling how close their behavior is to the norms of society.

Corollary 3. *When $\theta = 0$,*

$$s^t = \frac{\xi_1}{\xi_2 - \kappa_1} - \frac{\kappa_1}{\xi_2 - \kappa_1} p^t \tag{37}$$

That is, the probability of sharing toxic content only depends on user tastes for toxic content when there is no updating.

This results enable an interpretation of the reduced form results as the experiment does not warrant any exogenous variation in θ. The experiment provides insight into the effect of changing assignment probabilities exogenously.

D Proofs

Proof of Proposition 1

Consider a special case of the model so that sharing behavior does not respond to treatment assignment or the exposure to toxic content during the intervention period. Further the probability of sharing toxic content, for given levels of toxic exposure, is constant across all users. Then the proportionate change in sharing toxic content is exactly equal to the proportionate change in toxic content viewed, i.e. $ATE^{pShare} = ATE^{pView}$

Proof. We have potential outcomes $v_i^t(1)$ and $v_i^t(0)$ but s_i^t is unaffected by treatment or the realization of v_i^t under treatment assignment. Then, regression outcome S_i^t, the total number of toxic posts shared, on randomly assigned D_i identifies a slope coefficient

$$E[S_i^t(1) - S_i^t(0)] = mE[s_i^t \cdot (v_i^t(1) - v_i^t(0))] \equiv \gamma_1$$

and an intercept of

$$E[S_i^t(0)] = mE[s_i^t \cdot v_i^t(0)] \equiv \gamma_0$$

Consider also the regression of V_i^t, total number of toxic posts viewed, on randomly assigned treatment, D_i. This yields a slope coefficient of

$$E[V_i^t(1) - V_i^t(0)] = mE[v_i^t(1) - v_i^t(0)] \equiv \alpha_1$$

and an intercept of

$$E[V_i^t(0)] = mE[v_i^t(0)] \equiv \alpha_0$$

Then, we claim that $ATE_t^{pView} = ATE_t^{pShare}$ or $\gamma_1/\gamma_0 = \alpha_1/\alpha_0$. Consider,

$$\begin{aligned}
\frac{\gamma_1}{\gamma_0} &= \frac{mE[s_i^t(v_i^t(1) - v_i^t(0))]}{mE[s_i^t v_i^t(0)]} \\
&= \frac{mE\Big[(v_i^t(1) - v_i^t(0))E[s_i^t|v_i^t(0)]\Big]}{mE\Big[v_i^t(0)E[s_i^t|v_i^t(0)]\Big]} \\
&= \frac{m\pi \cdot E[v_i^t(1) - v_i^t(0)]}{m\pi \cdot E[v_i^t(0)]} = \frac{\alpha_1}{\alpha_0}
\end{aligned}$$

where the last two steps follow from monotonicity, non-responsiveness assumptions and the law of iterated expectations. □

Proof of Proposition 2

For $q_{i,\tau}^t = 1 - q_{i,\tau}^n$, user chooses total number of posts to view as

$$\nu_{i,\tau} = \left\{0, \frac{1}{2\eta}\left[1 + \alpha\theta - 2\alpha q_{i,\tau}^t q_{i,\tau}^n \theta + \alpha(1-\theta)\left(\left(p_i^t\right)^2 + \left(p_i^n\right)^2\right) - \frac{\left(q_{i,\tau}^t\right)^2}{\beta_i^t} - \frac{\left(q_{i,\tau}^n\right)^2}{\beta_i^n} - \frac{\alpha^2}{2}\right]\right\}$$

Proof. Substituting users' optimal strategy in the utility function, and measuring social preferences using the assignment probabilities generated by the algorithm gives,

$$u_{i,\tau} = \nu_{i,\tau}\left[q^t_{i,\tau}+q^n_{i,\tau}-\frac{\left(q^t_{i,\tau}\right)^2}{\beta^t_i}-\frac{\left(q^n_{i,\tau}\right)^2}{\beta^n_i}+\alpha\left[(1-\theta)\left(\left(p^t_i\right)^2+\left(p^n_i\right)^2\right)+\theta\left(\left(q^t_{i,\tau}\right)^2+\left(q^n_{i,\tau}\right)^2\right)\right]-\frac{\alpha^2}{2}-\eta\nu_{i,\tau}\right] \tag{38}$$

The claim follows from user's first order condition maximizing utility in (38) with respect to $\nu_{i,\tau}$. □

Proof of Proposition 4

In equilibrium, sharing probabilities are given by user tastes as well as the weight a user puts on their perception of society's preferences.

$$s^t_{i,\tau} = (1-\theta)p^t_i + \theta\frac{\left(\beta^t_i\right)^2}{(\beta^t_i+\beta^n_i)^2} + O(\theta^2)$$
$$s^n_{i,\tau} = (1-\theta)p^n_i + \theta\frac{\left(\beta^n_i\right)^2}{(\beta^t_i+\beta^n_i)^2} + O(\theta^2)$$

Proof. The proof follows from substituting the optimal assignment probabilities from the firm's problem in Lemma **??**, into users' best response functions described in Lemma 1. The final expression is achieved by applying a second degree Taylor approximation to linearize the expression around θ. □

Proof of Proposition 6

For some updating parameter θ and constant share of toxic posts λ, 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,

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

Proof. The equilibrium condition on users' sharing behavior in (25) simplifies to

$$s^t_{i,0} = \left(v^t_{i,0}\right)\left(w^t_i\right)^{\frac{\mu}{\theta}} \tag{39}$$
$$s^n_{i,0} = \left(v^n_{i,0}\right)\left(w^n_i\right)^{\frac{\mu}{\theta}} \tag{40}$$

As a result, the preference shock for toxic content is expressed as

$$w^t_i = \frac{w^n_i\left(\frac{s^t_{i,0}}{s^n_{i,0}}\right)^{\frac{\theta}{\mu}}}{\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right)^{\frac{\theta}{\mu}}} \tag{41}$$

Further, treated users view a constant proportion of toxic content during the intervention period $v^t_{i,1}(1) = \frac{\lambda \bar{q^t} v^n_{i,1}}{(1-\lambda)\bar{q^n}}$, because

$$v^t_{i,1}(1) = \frac{\lambda \bar{q^t}}{(1-\lambda\bar{q^n})\lambda\bar{q^t}}$$
$$v^n_{i,1}(1) = \frac{(1-\lambda)\bar{q^n}}{(1-\lambda\bar{q^n})\lambda\bar{q^t}}$$

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 (39) and (41) into the definition of the log-odds ratio of sharing toxic content in (23), so that

$$\log\left(\frac{s^t_{i,1}(1)}{s^n_{i,1}(1)}\right) = \log\left(\frac{\lambda^\theta \bar{q^t}^\theta \left(v^t_{i,0}\right)^{-\theta}\left(s^{}_{i,0}\right)}{(1-\lambda)^\theta \bar{q^n}^\theta \left(v^n_{i,0}\right)^{-\theta}\left(s^n_{i,0}\right)}\right) \tag{42}$$

where, $v^t_{i,1}$ is substituted from (**??**). Then, subtracting the defined log-odds ratio of sharing toxic posts (as defined above) at baseline ($\tau = 0$), and using (39) on the right hand side, we get

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

which is the required expression in the stated proposition. □

Proof of Proposition 7

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

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

(SA2) users engagement behavior in equilibrium is stable over time $s^t_{i,0} = s^t_{i,1}$ and $s^n_{i,0} = s^n_{i,1}$

(SA3) toxic posts form a fixed fraction of the corpus of content in each time period, λ

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

Then, θ can be estimated using the following regression equation

$$\mathrm{E}[\Delta B_{i,1}|D_i = 1] = \gamma_0 + \gamma_1 A_{i,0} \tag{43}$$

where, $\gamma_1 = -\theta$.

Proof. Consider Proposition 6 which gives

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

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,

$$\log\left(\frac{v^t_{i,0}}{v^n_{i,0}}\right) = \frac{1}{\kappa_1}A_{i,0}(1) + O(A_{i,0}(1))^2 \tag{44}$$

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, and is accounted for in the constant term. □

Proof of Proposition 8

Instrumenting average toxic views for each user with the average of one half of posts viewed removed bias in estimates on account of measurement error.

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

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

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

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

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

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

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

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

Proof of Proposition 11

When $s^t - (1-\theta)p^t \leq \theta q^t$, an intervention that randomly reduces q^t is costless to the firm.

Proof. The first order condition with respect to total number of posts viewed ν implies

$$\begin{aligned} \nu = \frac{1}{2\eta}\Bigg[& u(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\kappa_1}{\xi_2-\kappa_1}q^t - (1-\theta)\frac{\kappa_1}{\xi_2-\kappa_1}p^t; q^t, \beta^t) \\ & + u(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\kappa_1}{\xi_2-\kappa_1}q^n - (1-\theta)\frac{\kappa_1}{\xi_2-\kappa_1}p^n; q^n, \beta^n) \\ & - c(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\xi_2}{\xi_2-\kappa_1}q^t - (1-\theta)\frac{\xi_2}{\xi_2-\kappa_1}p^t) \\ & - c(\frac{\xi_1}{\xi_2-\kappa_1} - \theta\frac{\xi_2}{\xi_2-\kappa_1}q^n - (1-\theta)\frac{\xi_2}{\xi_2-\kappa_1}p^n)\Bigg] \end{aligned}$$

The proposition flows from the optimal ν function. □

Proof of Proposition 13

When $\theta = 0$, and for $\zeta_2^t(\beta^t), \zeta_2^n(\beta^n) \leq 0$

$$q^t = \frac{\zeta_1^n + \zeta_2^n - \zeta_1^t}{\zeta_2^t + \zeta_2^n} \tag{45}$$

That is, the probability of assigning toxic posts is directly proportional to users' distaste for non-toxic content, when there is no updating.

Proof. With $q^n = 1 - q^t$, algorithm's the first order condition implies

$$\begin{aligned} -\frac{\partial u(s^t; q^t, \beta^t, \alpha)}{\partial s^t}\frac{\theta\kappa_1}{\xi_2-\kappa_1} + \frac{\partial u(s^t; q^t, \beta^t, \alpha)}{\partial q^t} + \frac{\partial u(s^n; q^t, \beta^n, \alpha)}{\partial s^n}\frac{\theta\kappa_1}{\xi_2-\kappa_1} \\ - \frac{\partial u(s^n; q^n, \beta^n, \alpha)}{\partial q^n} + \frac{\partial c(s^t(q^t) - \theta q^t - (1-\theta)p^t)}{\partial q^t}\frac{\theta\xi_2}{\xi_2-\kappa_1} \\ - \frac{\partial c(s^n(q^n) - \theta q^n - (1-\theta)p^n)}{\partial q^n}\frac{\theta\xi_2}{\xi_2-\kappa_1} = 0 \end{aligned}$$

As before, for given assumptions on quadratic u and c (concave and convex, respectively), I write

$$\frac{\partial u(s^t; q^t, \beta^t, \alpha)}{\partial s^t} = \xi_1(\beta^t) + \xi_2(\beta^t)s^t \tag{46}$$

$$\frac{\partial u(s^n; q^n, \beta^n, \alpha)}{\partial s^n} = \xi_1(\beta^n) + \xi_2(\beta^n)s^n \tag{47}$$

$$\frac{\partial u(s^t; q^t, \beta^t, \alpha)}{\partial q^t} = \zeta_1^t(\beta^t) + \zeta_2^t(\beta^t)q^t \tag{48}$$

$$\frac{\partial u(s^n; q^n, \beta^n, \alpha)}{\partial q^n} = \zeta_1^n(\beta^n) + \zeta_2^n(\beta^n)q^n \tag{49}$$

$$\frac{\partial c(s^t(q^t) - \theta q^t - (1-\theta)p^t)}{\partial q^t} = \kappa_2(c)q^t \tag{50}$$

$$\frac{\partial c(s^n(q^n) - \theta q^n - (1-\theta)p^n)}{\partial q^n} = \kappa_2(c)q^n \tag{51}$$

where, $\xi_2, \zeta_2^t, \zeta_2^n < 0$ and $\kappa_2 > 0$. Solving for q^t, we get

$$\begin{aligned} q^t[2\theta^2\kappa_1^2 + (\zeta_2^t + \zeta_2^n)(\xi_2 - \kappa_1)^2 + 2\theta\xi_2\kappa_2(\xi_2 - \kappa_1)] &= \theta^2\kappa_1^2\xi_2 - \theta(1-\theta)\kappa_1^2\xi_2 p^n \\ &\quad + (\zeta_1^n + \zeta_2^n - \zeta_1^t)(\xi_2 - \kappa_1)^2 + \theta\xi_2\kappa_2 q^t(\xi_2 - \kappa_1) \end{aligned}$$

□

E Algorithms

Simulate Treatment

This algorithm simulates control and treatment embeddings over multiple days and analyzes the final aggregated embeddings. The treatment embeddings are generated by uniformly sampling from an epsilon ball, and the process is repeated for each user over several days. The Central Limit Theorem (CLT) suggests that the aggregated treatment embedding distribution should approximate a normal distribution.

Algorithm 1 Simulating Control and Treatment Embeddings

Input: $num_control_users$, $num_treatment_users$, num_movies, $embedding_dim$, num_days
Output: $final_control_embeddings$, $final_treatment_embeddings$
Generate synthetic binary user-item ratings data with missing values:

- Ratings are $\{1, 0, \text{NaN}\}$ with probabilities $\{0.3, 0.3, 0.4\}$

Remove rows with only NaN values
Calculate the global mean rating excluding NaN values
Fill in the missing values with the global mean
Apply Singular Value Decomposition (SVD) to derive embeddings:

- $U, \sigma, Vt = \text{svds}(ratings_filled, k = embedding_dim)$
- $initial_control_embeddings = U \cdot \sigma$

Initialize daily control and treatment embeddings arrays:

- $control_embeddings = \text{zeros}(num_control_users, embedding_dim, num_days)$
- $treatment_embeddings = \text{zeros}(num_treatment_users, embedding_dim, num_days)$
- $control_embeddings[:, :, 0] = initial_control_embeddings$

Calculate $control_mean$ and $control_var$ from $initial_control_embeddings$
Set $radius = 2 \cdot \sqrt{\sum control_var}$
for day in 0 to $num_days - 1$ **do**
  if day > 0 **then**
    Update control embeddings based on new "usage" with some noise
    $control_embeddings[:, :, day] = control_embeddings[:, :, day - 1] + \text{normal}(0, 0.1, \text{shape})$
  end if
  for user in 0 to $num_treatment_users - 1$ **do**
    while True **do**
      Generate a random point uniformly from the epsilon ball
      if $\text{norm}(point) \leq radius$ **then**
        $treatment_embeddings[user, :, day] = control_mean + point$
        break
      end if
    end while
  end for
end for
Aggregate the daily embeddings to get the final treatment and control embeddings:

- $final_treatment_embeddings = \text{mean}(treatment_embeddings, axis = 2)$
- $final_control_embeddings = \text{mean}(control_embeddings, axis = 2)$

Plot hexbin plots for the final aggregated embeddings
Plot histograms with proportions for each dimension with the same bin width
Calculate and print the correlation matrices for the final aggregated embeddings

Simulate Assignment Probabilities

Algorithm 2 Simulation and Derivative Calculation

Input: Constants α and k, range of θ, number of samples N
Output: Simulated values of $q_{i,\tau}^t$ and its derivative with respect to θ
Set $\alpha \leftarrow 1.0$
Set $k \leftarrow 0.5$ {Ensure $0 < k < 1$}
Generate N samples of β_i^t from a Beta distribution: $\beta_i^t \sim \text{Beta}(2, 2)$
Compute β_i^n as: $\beta_i^n \leftarrow \frac{k}{\beta_i^t}$
Clip β_i^n to the range $[0, 1]$
Define the range of θ values: $\theta \in [0.01, 1.0]$
for each β_i^t and β_i^n **do**
  for each θ value **do**
    Compute $q_{i,\tau}^t$ using the formula:

$$q_{i,\tau}^t = \frac{\beta_i^t(\beta_i^n\theta\alpha - 1)}{2\alpha\beta_i^t\beta_i^n\theta - \beta_i^t - \beta_i^n}$$

    Compute the numerical derivative of $q_{i,\tau}^t$ with respect to θ:

$$\frac{\partial q_{i,\tau}^t}{\partial\theta} \approx \frac{q(\theta + h) - q(\theta)}{h}$$

    Clip $q_{i,\tau}^t$ values to the range $[0, 1]$
  end for
end for
Plot the results:

- Plot $q_{i,\tau}^t$ against θ for different percentiles
- Plot the numerical derivative of $q_{i,\tau}^t$ against θ

Simulate Optimal Sharing Functions

Algorithm 3 Simulation and LOESS Smoothing for Sharing Functions

Input: Constants α, k, μ, intercept offset, range of θ, number of samples N
Output: Simulated values of $q_{i,\tau}^t$ and $s_{i,\tau}^t$ with LOESS smoothing
Set $\alpha \leftarrow 1.0$
Set $k \leftarrow 0.5$ {Ensure $0 < k < 1$}
Set $\mu \leftarrow 0.5$
Set intercept offset to 0.2
Define the range of θ values: $\theta \in \{0.1, 0.3, 0.5, 0.7, 0.9\}$
Generate N samples of β_i^t from a Beta distribution: $\beta_i^t \sim \text{Beta}(2, 2)$
Compute β_i^n as: $\beta_i^n \leftarrow \frac{k}{\beta_i^t}$
Clip β_i^n to the range $[0, 1]$
Generate p_i^t positively correlated with β_i^t
$p_i^t \leftarrow \min(1, \max(0, \beta_i^t + \text{Normal}(0, 0.1)))$
Generate w_i^t as a random normal draw centered around 0.5
$w_i^t \leftarrow \min(1, \max(0, \text{Normal}(0.5, 0.1)))$
for each θ value **do**
    Initialize arrays for $q_{i,\tau}^t$ and $s_{i,\tau}^t$
    for each sample **do**
        Compute $q_{i,\tau}^t$ using the formula:

$$q_{i,\tau}^t = \frac{\beta_i^t(\beta_i^n\theta\alpha - 1)}{2\alpha\beta_i^t\beta_i^n\theta - \beta_i^t - \beta_i^n}$$

        Ensure $q_{i,\tau}^t \geq 1e-5$ to avoid non-positive values
        Compute $s_{i,\tau}^t$ using the log-sharing function with offset:

$$s_{i,\tau}^t = \exp((1-\theta)\log p_i^t + \theta \log q_{i,\tau}^t + \mu \log w_i^t) + \text{offset}$$

    end for
    Perform LOESS smoothing on the $s_{i,\tau}^t$ values
    Plot the LOESS-smoothed curve
end for
Plot the 45-degree line for reference
Add labels, legends, and gridlines for better readability