Social Approval Seeking and Misperceived Legitimization: The Roots of Toxicity in Online Communities

Jun Goto Tsuyoshi Tsuru*

January 30, 2025

Abstract

Online toxicity has become a major global concern. However, we know little about how monetary and non-monetary incentives give rise to such toxicity and how online platform designs shape these incentives. This paper investigates whether mutual rating systems, illustrated by a "like" button, induce approval-seeking behavior and lead to the spread of toxicity. For this purpose, we draw on two natural experiments from a major Japanese Q&A platform: (i) the introduction of a like-button feature and (ii) the subsequent integration of these ratings with monetary incentives. Using a difference-in-differences approach combined with machine learning-based toxicity measures, we compare changes in toxicity levels before and after these interventions among users who are particularly sensitive to social approval versus those who are not. We find that individuals driven by social approval become significantly more toxic once these features are introduced, indicating that public affirmation can intensify harmful behaviors. Further analysis reveals a "misperceived legitimization" mechanism: approval-seeking users initially conceal their harmful tendencies, but interpret visible endorsements of toxic content as an indication of wider community acceptance, causing them to unleash their latent toxic behavior. These results suggest that platform features that promote social approval seeking can originate toxicity in online communities.

Keywords: approval seeking, toxicity, misperception

*Jun Goto (corresponding author): National Graduate Institute for Policy Studies, 7-22-1 Roppongi, Minato-ku, Tokyo 106-8677, Japan, ju-goto@grips.ac.jp; Tsuyoshi Tsuru: Hitotsubashi University, 2-1 Naka, Kunitachi, Tokyo 186-8603, Japan, tsuru@ier.hit-u.ac.jp; We thank Oriana Bandiera, Jed DeVaro, Takuma Kamada, Kohei Kawaguchi, Shuhei Kitamura, Eliana La Ferrara, David Levine, Ro'ee Levy, James Lincoln, Sultan Mehmood, Fumio Otake, Kazushi Takahashi, Kensuke Teshima, Norio Tokumaru, Stephen Vogel, Junichi Yamasaki, and Naoki Yoshihara as well as conference and seminar participants for helpful comments and suggestions.

1 Introduction

Online toxicity, including defamation, harassment, and hostile interactions, has become a significant global concern (Allcott et al., 2020a; Acemoglu et al., 2024). Therefore, understanding the roots of these behaviors is more urgent than ever. Recent literature explains the rise of toxicity from different perspectives: anonymity of online platforms can reduce accountability and heighten aggression (Ederer et al., 2024); network homophily fosters echo chambers that limit exposure to diverse perspectives, thereby radicalizing opinions (Sunstein, 2018; Cinelli et al., 2021; Levy, 2021); and individually tailored algorithms create filter bubbles that reinforce biases, further inflaming divisive discourse (Pariser, 2011; Guess et al., 2023b).

However, despite its long-standing recognition in the social sciences, one of the most important non-monetary incentives—*the desire for social approval*—has not been fully investigated.[1] Consequently, two crucial questions remain: (i) whether the pursuit of social approval causes harmful engagement in online communities, and (ii) whether platform designs that encourage approval seeking exacerbate toxic behavior.

This paper sheds light on the most prevalent design of online platforms: the mutual rating system, in which users rate each other's contributions to their online communities. Since Facebook introduced the "Like" button in February 2009, major platforms such as Twitter (now X) have continued to employ this design for more than a decade. By publicly displaying individual evaluations, mutual rating systems amplify users' desire for social approval (Hallinan and Brubaker, 2021). We therefore ask whether this emphasis on social approval, embedded in the mutual rating system, leads to the proliferation of hatred and toxicity online.

We focus on one of the largest online Q&A sites in Japan. This online platform provides an ideal setting for investigating the impacts of a mutual rating system on online toxicity because it experienced two sequential natural experiments. In the first experiment, the platform introduced a mutual rating system without prior notice. This system allowed users to send "tokens" to others as a way of expressing gratitude. Before its intro-

[1] Historically, the need for social approval has been regarded as one of the most important non-monetary incentives in the social sciences (Walther, 2022). The classic sociological and psychological literature has long recognized the human desire to gain acceptance, recognition, and validation from peers (Asch, 1951; Festinger, 1954; Baumeister and Leary, 1995). Recent research observes a close connection between the pursuit of social approval and online toxicity (Jiang et al., 2023); social approval seeking also drives online sharing behaviors by improving self-presentation (Nie et al., 2024). Indeed, a study by the Pew Research Center finds that adolescents are keenly aware of their presence on social media and the approval they receive from peers. See Anderson, M. & Jiang, J. (2018). *Teens, Social Media & Technology 2018. Pew Research Center.* (https://www.pewresearch.org/internet/2018/05/31/teens-social-media-technology-2018/)

duction, users had no indication of community approval levels. However, once the system was in place, everyone could see how many tokens each user and post had received. In other words, the new feature made social approval both visible and quantifiable.

In the second experiment, which occurred three months later, the platform allowed users to convert their accumulated tokens into monetary coupons. This added a financial dimension to the previously non-monetary "approval" currency. Taken together, the first natural experiment enables us to investigate whether introducing a mutual rating system lead the approval-seekers to post more harmful or toxic content, while the second experiment allows us to assess whether adding a monetary incentive to these visible ratings further influences harmful engagement.

We employ a difference-in-differences (DiD) approach to analyze how toxicity levels differ between two groups of users before and after specific platform interventions. Users are divided into two groups: those with a strong need for public approval (treatment group) and those with a minimal need for approval (control group). Based on established social psychology literature, we use the presence of public profiles on social media as a proxy for users' desire for social validation. Public profiles on platforms like Twitter increase visibility, enabling individuals to seek approval from wider audiences through likes, comments, and shares (Tang, 2022; Meythaler et al., 2023; Sciara et al., 2023). In contrast, users without social media accounts or with private profiles (e.g., protected accounts on Twitter) have reduced visibility, limiting opportunities for public endorsement (Gruzd and Hernández-García, 2018).[2]

To implement this concept, we identify public profiles by checking whether users on our platform have social media accounts (using the same username) and whether those profiles are public. By combining this measure with our DiD approach, we examine whether users with public profiles—indicating a stronger need for social approval—respond more strongly to interventions such as the introduction of a like button or monetary incentives, reflecting their emphasis on public validation.

In measuring toxicity, we focus exclusively on users' answers rather than on the questions. Note that the design changes may influence both respondents and question posters. Therefore, it is not clear whether any observed shifts in answers are due to respondents' own behavior or to their reactions to more provocative questions. To isolate respondents' behavior, we take advantage of the platform's 15-day limit for submitting answers: once a question is posted, responses can only be added for 15 days. More precisely, we use a

[2]These distinctions are supported by Social Comparison Theory (Festinger, 1957; Suls and Wheeler, 2012) and Uses and Gratifications Theory (Katz et al., 1973; Cipolletta et al., 2020), which suggest that people seeking approval tend to choose public profiles to engage in social comparison and derive self-esteem from visible feedback metrics.

subsample of answers that are posted for questions registered before the system changes but still open after those changes took effect. This enables us to compare answers within the same threads before and after the design updates while keeping the question content constant. This ensures that any observed changes in toxicity are driven by respondents' behavior rather than by differences in the questions themselves.

We quantify the toxicity by focusing on linguistic nuances and emergent internet slang. Specifically, we use FastText, a word-embedding model that accommodates subword units, allowing us to identify evolving patterns of harmful language in Japanese. First, we train the FastText model on a unique dataset that includes both regular posts and those removed by moderators due to ethical violation. Second, using these removed toxic posts as a semantic anchor, we produce a "toxicity score" for each word, reflecting its relative proximity to this semantic anchor (i.e., a reference of harmful language) within the embedding space. This allows us to define a toxic–nontoxic continuum, projecting words onto a scale that captures subtle variations in language use. Finally, we aggregate these word-level scores at the document level to generate our primary outcome variable.

We obtain two main findings. First, in the first natural experiment, which introduces a like-button system that makes social approval visible to all, we find a rapid escalation in toxic content among users who value public recognition. More precisely, our estimates indicate that approval-seeking users increase their toxicity by approximately 0.22 standard deviations after the platform redesign. It suggests that displaying social approval encourages certain individuals to shift towards more ethically problematic discourse, amplifying the intensity of harmful expressions.

Second, turning to another natural experiment, where the platform ties user ratings to monetary incentives, we observe an even more pronounced increase in toxicity: approval-seeking users intensify their toxicity by about 0.49 standard deviations relative to non-approval-seeking users. This substantial effect demonstrates the role of monetary incentives in fueling extreme and harmful viewpoints: individuals who are inclined to seek approval push the boundaries of acceptable communication even further.

We address threats to internal validity based on three approaches. First, we implement event-study analysis and confirm that the parallel trends assumption hold in our context. Second, we examine spillover effects to test the validity of the Stable Unit Treatment Value Assumption (SUTVA). The results show no spillovers: control users do not become more toxic even when exposed to higher shares of treated users or their toxic content. This ensures that our comparison group remains a valid counterfactual. Finally, we verify that the observed changes are not driven by shifts in user composition. We find no post-treatment selection effects: the proportion and characteristics of treated versus control

users remain stable after the interventions.

Next, we test whether our main findings are indeed driven by users' desire for social approval. Our analysis rejects the alternative explanation. First, by distinguishing between "active" and "inactive" public account holders, we find that only active users—those who actively produce content on social media—show increased toxicity following the interventions; inactive users with public profiles exhibit no such change. Second, we identify non-anonymous users, defined as those who disclose identifiable details like job titles or affiliations, and observe that their toxicity levels remain unchanged. Together, these results suggest that the increase in toxicity among public-account holders is driven by their desire for social approval, not by other intrinsic motivations associated with having public accounts and their non-anonymous status on our platform.

We turn to investigate the heterogeneity in our main findings by identifying which users are most affected by the design changes. Our analysis shows that users who are both "potentially harmful" and "previously hidden" are especially likely to escalate their behavior once the new system designs take effect. Specifically, we define "potentially harmful" individuals as those whose language patterns resemble those of users who have previously violated ethical rules. Additionally, users are "previously hidden" if they have never had a post removed for ethical violations before the intervention. Our analysis reveals that these users more prominently increase toxicity once the new system designs are introduced.

We also examine whether toxic content actually attracts more support from other users. Interestingly, while public-profile users who become more toxic do receive additional "likes" from other toxic users, they ultimately lose approval from the wider user community, resulting in fewer overall "likes." This finding highlights a misperception: toxic behavior does not offer a reliable path to broad popularity. Instead, it only resonates with similarly toxic audiences and fails to generate any net gains in support.

Then, why do potentially harmful yet previously hidden users respond more aggressively to new system designs? Our analysis suggests that these users misinterpret "liked" toxic comments displayed on the top page as signals of community-wide approval. Observing toxic posts that receive visible endorsements encourages them to intensify their harmful behavior, hoping to gain attention and validation. Although this strategy does garner some positive feedback from similarly toxic users, it alienates the broader community.

We refer to this phenomenon as *misperceived legitimization*, where users with latent toxic tendencies, who conceal them before system changes, interpret "liked" toxic comments as signals that harsher language is not only acceptable but that their underlying

toxicity is validated. To examine this mechanism, we use a DiD approach across three scenarios: (1) when no toxic posts appear on the top page, (2) when toxic posts without likes appear on the top page, and (3) when toxic posts with likes appear on the top page. Our findings show a significant increase in toxicity only in the third scenario, suggesting that visible reinforcement—rather than the mere presence of toxic content—motivates harmful users to escalate their behavior. This indicates that such users misinterpret approval from a vocal minority as broad societal acceptance, leading them to feel legitimized in expressing their inherent but previously hidden hatred.

Finally, we examine how rising toxic sentiments contribute to the formation of segregated communities, or "echo chambers." We focus on a Q&A platform where users can "follow" others, allowing us to track friendships and measure toxicity scores between pairs of users. We construct a dyadic, monthly panel dataset—each pair forms a single observation—and employ a simple OLS model to test whether friendships established in a given month exhibit greater toxicity similarity after the system changes. Specifically, we use the difference in toxicity scores between user pairs as our outcome and include an interaction term between a "friend registration" dummy (one if a pair becomes friends in that month) and a "post-system design change" dummy (one for periods following the interventions). Our findings show that newly formed friendships increasingly involve users with similar toxicity levels. This clustering effect indicates that the revised incentive structures encourage toxic users to connect, leading to fragmented, like-minded communities that reinforce and heighten toxic behavior.

Related Literature This paper contributes to three strands of literature. First, it relates to the active body of research exploring how non-monetary incentives influence decision-making in offline communities and their interaction with monetary incentives (Bowles, 2008; Bowles and Polania-Reyes, 2012; Ashraf and Bandiera, 2018; Besley and Ghatak, 2018). Non-monetary incentives are also widely recognized as crucial drivers of content production on social media platforms (Ma, 2023; Aridor et al., 2024). While this growing field highlights the significant role of these incentives in producing content, there is limited evidence on how the social approval-seeking preference fosters toxicity in online communities or how it interacts with monetary incentives.[3] Our study addresses this gap

[3] Aridor et al. (2024) identify five key non-monetary incentives: (i) attention and visibility (Abreu and Jeon, 2019; Acemoglu et al., 2023), (ii) enhanced social image and reputation (Chen et al., 2010; Bohren et al., 2019; Filippas et al., 2022), (iii) feedback and peer rewards (Eckles et al., 2016; Zeng et al., 2023), (iv) persuasion and advocacy (Bursztyn et al., 2023), and (v) intrinsic or altruistic motives (Guriev et al., 2023). These incentives propagate through networks, creating short-lived cascading effects (Huang and Narayanan, 2020; Mummalaneni et al., 2023; Eckles et al., 2016; Zeng et al., 2023), and even negative feedback like downvotes can spur additional content creation (Deolankar et al., 2023). Peer networks am-

by providing the first evidence on this question, leveraging unique natural experiments in Japan: the introduction of a like-button system and rating-based monetary incentives. Relatedly, this paper also intersects with the literature on motivational crowding out, which explores how monetary incentives can undermine intrinsic motivation (Bowles, 2008; Gneezy et al., 2009; Bowles and Polania-Reyes, 2012). We reveal the unintended effects of monetary incentives on online platforms, where platform designs amplify social approval-seeking preferences, ultimately discouraging pro-social behaviors.

This paper also speaks to the emerging literature on polarization, discrimination, and hate speech in online communities (Zhuravskaya et al., 2020; Campante et al., 2023; González-Bailón and Lelkes, 2023). Recent papers in economics and other social sciences find that Internet and social media usage may facilitate political polarization, societal segregation, extreme ideology, hate crimes, and political protests (Allcott et al., 2020a; Bursztyn et al., 2019; Lelkes et al., 2017; Mosquera et al., 2020; Levy, 2021; Müller and Schwarz, 2021,0).[4] Our paper complements this literature by suggesting a new mechanism through which online hatred and toxicity emerge: The mutual rating system, which is widely used on online platforms, emphasizes the need for social approval, leading to extremism and segregated communities. Our evidence is unique because it explains why toxicity arises on many online platforms: Users' rational responses to social approval incentives result in more radicalized statements, and unless additional institutional changes are undertaken, this toxicity might persist for an extended period of time.

This paper is also related to the literature on misperceptions that highlights how incorrect beliefs about others influence attitudes and behaviors, often driven by stereotypes,

plify these effects (Zeng et al., 2023), but perceived quality remains crucial, with modest improvements often assessed by likes (Zeng et al., 2023; Srinivasan, 2023). In contrast, monetary incentives strongly influence content creation; changes in revenue-sharing programs lead to sharp shifts in production (Kerkhof, 2024; El-Komboz et al., 2023). Social media also influences journalism by increasing story visibility and prompting headline revisions (Cagé et al., 2022; Leung and Strumpf, 2023), highlighting the complex interplay of incentives (Burtch et al., 2022; Srinivasan, 2023).

[4]It is worth emphasizing that there is mixed evidence about whether using the Internet leads to social polarization and extremism, and thus more discussion is needed to understand this better. On one hand, Boxell et al. (2017) note that US polarization is most pronounced among those who rarely use the Internet, and Gentzkow and Shapiro (2011) find that online interactions are less segregated than offline ones. On the other hand, Halberstam and Knight (2016) observe that communication on social media (e.g., Twitter) is highly segregated, mirroring offline patterns. Recent work indicates complexity. Guess et al. (2023a) argue that echo chambers' effect on polarization depends on pre-existing political attitudes, and Allcott et al. (2020b) find that exposure to misinformation correlates with partisan differences but does not always increase polarization. Bail et al. (2021) show that exposing users to opposing views often backfires, reinforcing established positions. Zhuravskaya et al. (2020) review how the Internet and social media shape political preferences and polarization. Prior (2013) find no conclusive evidence linking biased media to polarization. DiMaggio et al. (2001) highlight conflicting sociological research on Internet usage and social movements.

motivated reasoning, and pluralistic ignorance. Misperceptions are pervasive, larger for out-groups, and correlated with individual attitudes (Bursztyn and Yang, 2022). Studies, such as Bordalo et al. (2016) on stereotypes and Kuran (1997) on pluralistic ignorance, emphasize their persistence. Experimental work shows that accurate information can mitigate misperceptions, though with varying durability (Cantoni et al., 2019; Allcott et al., 2020a; Bursztyn et al., 2020). We contribute to this literature by introducing the concept of "misperceived legitimization" as a mechanism driving online toxicity. This mechanism posits that visible endorsements of harmful content, such as likes on toxic comments, create the false impression of societal acceptance, particularly among approval-seeking individuals. By identifying this mechanism, we extend the understanding of how platform design influences misperceptions and behavior. Unlike existing work, which primarily addresses static misperceptions, our findings reveal their roots where misperceptions emerge in response to design changes. This provides new evidence on the role of digital environments in shaping harmful social norms.

2 Natural Experiments

Our identification strategy relies on two sequential natural experiments: (i) the introduction of the mutual rating system, which makes social approval visible; and (ii) the provision of monetary incentives proportional to individual ratings. This section provides a detailed explanation of these natural experiments.

2.1 Context

General Information Our target Q&A site is one of the largest online platforms in Japan. This Q&A site records more than 80 million page views per month on average. During the period from 2017 to 2019, averaged 264.06 entries posted per day, with 94.2% of questions receiving at least one answer within a week. Our Q&A site shares common features, such as user-generated content, anonymous communication, and unrestricted information accessibility, with other popular platforms throughout the world (see Table A.1). Therefore, institutional features and communication procedures among users on this website are comparable with other representative Q&A sites.

On this website, any registered user can post a question at any time, and anyone can respond. Right after a question is posted, the platform administrator sets up a thread for it. All internet users, even those not registered, can view these threads. However, only users who are logged in can answer a question, and this is only within 15 days of it

being posted. Once the 15 days are up, the administrator closes the thread, and no more answers can be added. All threads are then stored in a public archive, which anyone can access at any time.

Concerns for Online Hatred and Extremism Like other major social media platforms, this Q&A website faces persistent challenges with online abuse and defamation. To address these issues, a basic algorithm continuously scans user-submitted questions and answers for forbidden content, compiles them into a list, and removes them—pending staff review and approval—without notifying users. Additionally, staff review and remove posts flagged by users as inappropriate. Between September 2017 and September 2019, the site eliminated 11.74% of all posts in accordance with its guidelines, which explicitly state that administrators may delete any content without warning if it risks causing (i) privacy breaches, (ii) copyright infringement, (iii) slanderous defamation, (iv) violations of public policy, or (v) any form of harassment.

The User Account Page User account pages on this Japanese QA platform are designed to give each member a customizable presence while keeping the mandatory requirements to a minimum. Every user has an account page accessible by clicking on their name, where optional fields such as handle name, age, residential prefecture, gender, preferred topics, and registration date can be displayed. Users can also include links to their social media profiles (e.g., Twitter, Instagram, TikTok, Facebook), though only a user account name is strictly required to sign up. Notably, 42.2% of users choose the same account name across this platform and other SNSs, and 17.2% openly share their SNS URLs. This high degree of social media integration stems from the convenient registration process, which allows new users to create an account simply by linking an existing social media profile—no extra steps required. In this way, user account pages not only serve as personal profiles but also facilitate easy cross-platform networking and interaction, reflecting the platform's emphasis on accessibility and community engagement.

2.2 The Introduction of the Mutual Rating System

Our Q&A site launched a new token system to promote user interaction and frequent postings. This system lets users rate each other's posts, effectively making social approval visible within the platform. Essentially, this token system is akin to the "like" button feature on other social media platforms, as it enables users to evaluate one another and publicly display individual ratings. This section outlines the specifics of this token system.

The token system is an integral part of the platform's user engagement and feedback mechanism. When it was first introduced, every user received thirty-nine tokens from the administrator. Users can distribute these tokens to any question or answer they find valuable, and they can do so regardless of whether they have posted in that thread themselves. Every Monday, the administrator resets each user's token count to thirty-nine, rendering any leftovers from the previous week invalid. While there is no limit to the number of tokens a user can give away (as long as they have them), tokens received cannot be passed on further. The number of tokens each post has earned—and the username of its poster—are always displayed just below the content, allowing everyone to see how the community rates both the post and its contributor. Before this token system was implemented, users had no way to gauge the public reception of individual posts, making the current setup a marked shift toward greater transparency and community-driven validation.

The token system was launched on June 26, 2018, at 12 p.m. From the start, tokens could neither be converted into Japanese yen nor used as coupons for purchasing other goods. The official explanation on the website did not mention any possibility of exchanging tokens for cash or equivalent value; instead, it highlighted their role in promoting public recognition of each user. Consequently, we believe this approach was unlikely to be regarded as introducing monetary incentives. Rather, it reinforced the idea that everyone on the platform could acknowledge one another's evaluations.

2.3 Provision of Monetary Rewards Proportional to Ratings

At 10 a.m. on September 20, 2018, the platform administrator announced a new feature enabling users to exchange their tokens for redeemable coupons provided by partnering private companies in Japan. On average, 12.8 coupons are available at any given time, each expiring in roughly 30 days. Users only see coupons currently on offer and receive no advance notice about upcoming coupons. Between 2017 and 2019, the exchange rate for tokens was 6.76 yen per token. The highest number of tokens awarded to a single post during this period was 345. Notably, 62.32% of posts received no tokens while their discussion threads remained open. On a weekly basis, users dispensed an average of 21.2 tokens. Moreover, 52% of users spent all the tokens issued to them before they expired.

2.4 Descriptive Statistics

The target Q&A site exhibits high user engagement, with an average of 1,285 users logging in per hour and spending approximately 43 minutes browsing. On average, new questions

receive their first answer within about 18.45 minutes. Among all questions, 98.2% receive at least one answer within a week, and 71.4% receive more than two. This high level of activity makes the site suitable for investigating how mutual rating systems affect user toxicity. Table A.2 presents summary statistics of user characteristics. Panel A corresponds to the sample in the first treatment, while Panel B corresponds to the sample in the second treatment. The table includes data on user demographics such as age and the length of site registration (in months). Overall, no significant differences emerge between the first and second treatment samples, nor between pre- and post-treatment samples.

3 Measuring Toxicity of User Posts

To investigate the impact of introducing the like button and rating-based monetary incentives on toxicity within the Q&A platform, it is crucial to accurately quantify the toxicity of each post using textual data. We employ a machine learning-based word embedding approach called FastText to achieve this task. In this section, we provide a detailed explanation of this method and elaborate on how we construct our target outcomes.

3.1 Textual Data

General Information In collaboration with the operator of the Q&A website, we collected complete textual data containing all questions and their corresponding answers posted between 2017 and 2019. This dataset consists of 190,612 questions and 724,925 answers. Additionally, we obtained data on answers that were deleted by platform moderators due to violations of community guidelines, particularly those related to ethical or behavioral standards. These deleted answers, totaling 15,342 posts, serve as references to toxic content and are used to develop our toxicity measurement model.

We focus on general toxicity in user posts rather than context-dependent hatefulness or extremism targeting specific social groups. Therefore, we use all available answers in our dataset, which spans various topics, including social, political, and economic issues. This broad range of topics helps us capture overall harmful content that is not tied to specific subjects or groups and includes any form of discriminatory or defamatory language used on the platform.

Text Pre-processing Following standard procedures in quantitative text analysis, we pre-processed the collected postings to prepare them for analysis. We tokenized the Japanese texts into individual words. To enhance tokenization, we leveraged collocation

analysis to identify words with strong associations and compound them, forming selective n-grams. To reduce noise in the data, we employed the widely accepted stop-words list for Japanese (i.e., Marimo), an extension of the Snowball stop-words list, to remove frequently used but less meaningful words. Additionally, we excluded words that occurred less than 10 times throughout the entire corpus to prevent the model from being influenced by extremely rare keywords.

3.2 Definition of Toxicity

We define toxicity in user posts as language that is rude, disrespectful, or unreasonable. To quantify the level of toxic language in posts, we consider two main approaches.

The Dummy for Removal by the Platform The first is a straightforward method that relies on whether a post is removed by the platform for violating community guidelines. The platform's guidelines expressly state that the management may remove any postings without notice if they could lead to (i) defamation through slander, (ii) a breach of public policy, or (iii) various forms of harassment. Because our dataset includes both publicly viewable posts and deleted posts, this approach allows us to evaluate how platform design changes influence the number of deleted posts.

However, this approach has several limitations: (a) It fails to capture harmful content when a post includes slanderous or discriminatory language that does not meet the threshold for removal. (b) The binary nature of deletion (deleted or not deleted) obscures more nuanced variations in the degree of defamation or harm caused by posts. (c) It is highly influenced by the platform's changing preferences or biases. For example, two posts with identical content could be labeled as toxic and non-toxic depending on shifts in the platform's standards. (d) It does not account for the context and nuance of entire sentences. Even if a post avoids using explicitly discriminatory keywords, it can still carry the intent to harm or demean others.

Semantic Similarity to Referenced Toxic Posts As a more precise and preferable second approach to measuring toxicity, we employ machine learning techniques. Intuitively, we assess the semantic similarity between each post and reference posts that are removed by the platform before the system design changes: If the linguistic features of a post are similar to those of posts that were removed for violating ethical rules, the post can be considered more likely to contain harmful content. Importantly, this measurement method, which takes into account the linguistic features and structure of all posts, ad-

dresses all of the aforementioned issues associated with using a binary variable to indicate whether a post was deleted. Hence, we adopt this machine-learning-based measurement approach as the primary method for assessing the harmfulness of post content.

3.3 The Machine Learning Model for Measuring Semantic Similarity

To measure the similarity between posts, we rely on the word embedding method. Word embedding is a technique that represents words as dense, continuous vectors in a high-dimensional space, where the distance between vectors reflects the semantic similarity between the words. This allows words with similar meanings to be positioned closer together, enabling meaningful comparisons based on their contextual relationships. We adopt the recently developed algorithm, i.e., FastText (Bojanowski et al., 2017; Joulin et al., 2017; Grave et al., 2018). FastText extends the traditional Word2Vec model by incorporating subword information through character n-grams. Unlike models that treat words as indivisible units, FastText represents each word as a collection of its character-level components. This approach enables the model to capture morphological nuances and effectively handle out-of-vocabulary words, which is particularly beneficial for languages with rich morphological structures or when dealing with rare or misspelled words.

3.3.1 The Skip-Gram Model with Negative Sampling

One of the key components that make up FastText is a technique called Skip-Gram with Negative Sampling (SGNS). SGNS is a method used to train word embeddings by predicting the context words surrounding a target word within a defined window size. Instead of using a computationally expensive full softmax for all possible words in the vocabulary, SGNS simplifies the process by employing negative sampling, where only a small subset of negative examples (words not in the context) is used during training. This approach significantly reduces the computational burden and serves as the foundational architecture upon which FastText builds its extended capabilities, such as subword representations.

Objective Function Given a corpus represented as a sequence of words

$$\underbrace{w_1, w_2, \ldots, w_T}_{\text{sequence of } T \text{ words}},$$

the Skip-Gram model aims to learn word embeddings by predicting the context words surrounding a target word. Here, T denotes the total number of words in the corpus,

and w_t is the target word at position t in the corpus. The words w_{t+j} are the context words within a window size c around w_t. The index j ranges from $-c$ to $+c$, excluding 0, indicating how far from the target word w_t we look for context words. Thus, for each target word w_t, the model attempts to predict the context words w_{t+j} within a window size c, where $-c \leq j \leq c$ and $j \neq 0$. The objective is to maximize the following log-likelihood function:

$$\mathcal{L} = \sum_{t=1}^{T} \sum_{\substack{-c \leq j \leq c \\ j \neq 0}} \log P(w_{t+j} \mid w_t).$$

Here, $P(w_{t+j} \mid w_t)$ is the conditional probability of observing the context word w_{t+j} given the target word w_t. However, computing the conditional probability $P(w_{t+j} \mid w_t)$ over the entire vocabulary is computationally intensive due to the large size of the vocabulary in natural language processing tasks.

Negative Sampling Negative sampling addresses this computational challenge by reframing the problem as a binary classification task. Instead of considering all possible context words for every target word, the model distinguishes between observed word-context pairs and randomly sampled noise pairs. An observed pair (w_i, w_o) is a true target-context pair actually present in the corpus, whereas noise pairs (w_i, w_k) are negative examples sampled from a noise distribution $P_n(w)$. These pairs do not co-occur in the context window for w_i. The parameter K is the number of negative samples (noise words) drawn for each observed pair. The goal is to assign a high probability to the observed pair (w_i, w_o) while assigning low probabilities to the negative (noise) pairs (w_i, w_k). The objective function for a single observed pair then becomes:

$$\mathcal{L}_{w_i, w_o} = \log \sigma \left(\mathbf{v}'^{\top}_{w_o} \mathbf{v}_{w_i} \right) + \sum_{k=1}^{K} \mathbb{E}_{w_k \sim P_n(w)} \left[\log \sigma \left(-\mathbf{v}'^{\top}_{w_k} \mathbf{v}_{w_i} \right) \right],$$

where $\sigma(x) = \frac{1}{1+e^{-x}}$ is the sigmoid function, $\mathbf{v}_{w_i}$ is the embedding vector for the input (target) word w_i, and $\mathbf{v}'_{w_o}$ and $\mathbf{v}'_{w_k}$ are the embedding vectors for the output (context) word w_o and noise words w_k, respectively. The distribution $P_n(w)$ is the noise distribution from which negative words w_k are sampled, and $\mathbb{E}_{w_k \sim P_n(w)}[\cdot]$ indicates the expected value under this noise distribution. Intuitively, the term $\log \sigma(\mathbf{v}'^{\top}_{w_o} \mathbf{v}_{w_i})$ increases when $\mathbf{v}'_{w_o}$ and $\mathbf{v}_{w_i}$ are similar (i.e., when the dot product is large), making it more likely for the model to recognize w_o as a valid context word for w_i. Meanwhile, the negative sampling term $\sum_{k=1}^{K} \log \sigma(-\mathbf{v}'^{\top}_{w_k} \mathbf{v}_{w_i})$ decreases when any randomly sampled word w_k is mistakenly considered a valid context, thereby penalizing associations with incorrect contexts.

Intuitive Explanation The objective function for negative sampling intuitively aims to maximize the model's ability to distinguish between "true" word-context pairs and randomly sampled "noise" pairs. For observed, or "true," word-context pairs (w_i, w_o), the term $\log \sigma\left(\mathbf{v}'^{\top}_{w_o}\mathbf{v}_{w_i}\right)$ reflects the probability that the context word w_o is genuinely associated with the target word w_i. The model seeks to maximize this probability by increasing the similarity between the vector representations of w_i and w_o, thereby making their dot product larger. This optimization encourages the sigmoid function $\sigma(\cdot)$ to approach 1, indicating a high likelihood that (w_i, w_o) is a valid pair.

For noise pairs (w_i, w_k), which are randomly sampled from a noise distribution, the objective includes a penalty term $\sum_{k=1}^{K} \mathbb{E}_{w_k \sim P_n(w)} \left[\log \sigma\left(-\mathbf{v}'^{\top}_{w_k}\mathbf{v}_{w_i}\right)\right]$. This term pushes the model to reduce the similarity between w_i and the noise words w_k, ensuring that their dot product is small and negative. By doing so, the sigmoid function $\sigma(-\cdot)$ is maximized, approaching 1 for these noise pairs, which represents a low probability of being true pairs.

Overall, the model balances these two goals: maximizing the probability for true pairs while minimizing it for noise pairs. This process encourages the learned word embeddings to position semantically related words close to one another in the vector space while keeping unrelated words far apart. Framed as a binary classification task, the model learns to assign high probabilities to positive (true) examples and low probabilities to negative (noise) examples, resulting in embeddings that effectively capture semantic relationships.

3.3.2 Incorporating Subword Information in FastText

Word Representation with Subword Units FastText enhances the Skip-Gram model by incorporating subword information, which means it breaks each word down into smaller units, such as prefixes, suffixes, or overlapping sequences of characters called n-grams. By representing words as collections of subword units, FastText captures internal structures and shared patterns among related words. For example, when the word "economics" is enclosed by boundary symbols as ⟨economics⟩, the model extracts various n-grams (for instance, ⟨*ec*, *eco*, *con*, and *ics*), each of which has its own vector representation. Words that share many of these n-grams, like "economy" or "economic," end up with similar vector representations. This design also helps when encountering rare or misspelled words, since unfamiliar words often still share certain subword units with known words.

Modified Objective Function The input vector for a word w_i in FastText is formed by summing the vector representations of its n-grams, as shown in the equation

$$\mathbf{v}_{w_i} = \sum_{g \in G_{w_i}} \mathbf{z}_g.$$

In this expression, $\mathbf{v}_{w_i}$ is the embedding vector for the word w_i. The symbol G_{w_i} denotes the set of all n-grams contained within w_i, and each n-gram g has an associated vector $\mathbf{z}_g$. The sum of these subword vectors serves as the overall representation of w_i. Meanwhile, the output word vectors $\mathbf{v}'_{w_o}$ remain as in the original Skip-Gram model, corresponding to the representations of context (output) words. The negative sampling objective function from the Skip-Gram model then incorporates these summed subword vectors:

$$\mathcal{L}_{w_i,w_o} = \log \sigma\Big(\mathbf{v}'^{\top}_{w_o}\big(\sum_{g \in G_{w_i}} \mathbf{z}_g\big)\Big) + \sum_{k=1}^{K} \mathbb{E}_{w_k \sim P_n(w)}\Big[\log \sigma\Big(-\mathbf{v}'^{\top}_{w_k}\big(\sum_{g \in G_{w_i}} \mathbf{z}_g\big)\Big)\Big].$$

Here, $\mathbf{v}'_{w_o}$ is the embedding vector for the context word w_o. The term $\sum_{g \in G_{w_i}} \mathbf{z}_g$ is the subword-based representation of the target word w_i. The notation $\sigma(\cdot)$ refers to the sigmoid function, while the summation $\sum_{k=1}^{K} \mathbb{E}_{w_k \sim P_n(w)}[\dots]$ implements negative sampling by drawing K noise words w_k from a noise distribution $P_n(w)$ and encouraging the model to assign them low probabilities of being correct contexts.

Training Objective The overall training objective is to maximize the sum of these individual loss terms over all observed word-context pairs in the dataset. If D represents the set of all such pairs, then the total objective is:

$$\mathcal{L} = \sum_{(w_i,w_o) \in D} \mathcal{L}_{w_i,w_o}.$$

The notation $(w_i, w_o) \in D$ indicates that for each target word w_i, there is an associated context word w_o in the corpus. By summing the loss terms across all observed pairs, FastText learns vectors for both n-grams and whole words that align well with the patterns present in actual text.

3.3.3 Application to Measuring Toxicity

Model Training FastText embeddings can be leveraged to detect toxicity in user-generated content by exploiting the morphological nuances captured by subword repre-

sentations. The model is trained on a corpus of user posts, learning embedding vectors for each n-gram and word found in the data. Since subword units help capture common patterns even across rare or novel words, this approach is well-suited for the varied and often informal language seen in online forums.

Constructing Post Vectors After training, each post p in the dataset can be represented as a vector $\mathbf{v}_p$ by averaging the subword-based embeddings of its constituent words:

$$\mathbf{v}_p = \frac{1}{|W_p|} \sum_{w \in W_p} \Big(\sum_{g \in G_w} \mathbf{z}_g \Big).$$

In this formula, W_p is the set of words in the post p. Each word w is itself represented by the sum of its n-gram vectors $\sum_{g \in G_w} \mathbf{z}_g$. Dividing by $|W_p|$ yields the average embedding for the entire post.

Creating a Referenced Toxicity Vector A referenced toxicity vector $\mathbf{v}_{\text{toxic}}$ is constructed by averaging the vectors of posts known to be toxic. If the set of deleted posts (due to ethical or moderation violations) is denoted by D, then

$$\mathbf{v}_{\text{toxic}} = \frac{1}{|D|} \sum_{p \in D} \mathbf{v}_p.$$

Here, each deleted post p in D is first converted into its averaged embedding $\mathbf{v}_p$, and then these vectors are aggregated to create a single reference point for toxicity.

Calculating Toxicity Scores A new post p receives a toxicity score by measuring its similarity to $\mathbf{v}_{\text{toxic}}$. This is done through the cosine similarity:

$$\text{Toxicity Score}(p) = \cos\Big(\mathbf{v}_p, \mathbf{v}_{\text{toxic}}\Big) = \frac{\mathbf{v}_p \cdot \mathbf{v}_{\text{toxic}}}{\|\mathbf{v}_p\| \|\mathbf{v}_{\text{toxic}}\|}.$$

The notation $\mathbf{v}_p \cdot \mathbf{v}_{\text{toxic}}$ denotes the dot product between the post vector and the toxicity vector, while $\|\mathbf{v}_p\|$ and $\|\mathbf{v}_{\text{toxic}}\|$ are their respective magnitudes (Euclidean norms). A higher cosine similarity indicates a greater semantic resemblance to posts deemed toxic, thus suggesting the presence of similar toxic language patterns.

3.3.4 Advantages of Using FastText

FastText offers several advantages by integrating character n-grams into its architecture, enabling it to capture both the semantic and morphological aspects of language. This

feature is particularly beneficial when working with Japanese, a language characterized by a complex writing system that includes kanji, hiragana, and katakana scripts. Japanese words often consist of intricate combinations of these scripts, along with prefixes, suffixes, and root words, making traditional word-level representations less effective.

In the context of a Japanese QA site, where unique words and new internet slang are prevalent, FastText excels by learning patterns within subword components. Online communities frequently generate neologisms and community-specific terminology that may not appear in standard vocabularies. Traditional models struggle with these out-of-vocabulary words, leading to gaps in understanding and representation.

By utilizing subword information through character n-grams, FastText can produce meaningful embeddings for rare or unseen words, including newly coined internet terms. This capability mitigates the problem of limited vocabulary and enhances the model's ability to comprehend and process user-generated content rich in unique expressions. Consequently, FastText is well-suited for applications involving dynamic and evolving language use, such as that found on Japanese QA platforms.

Distribution of Toxicity Scores We analyze the distribution of the toxicity scores to understand the prevalence and intensity of toxic content on the platform. Figure A.1 plots individual words according to two dimensions: their "toxicity" score (on the x-axis) and their frequency (on the y-axis, represented in log scale). Each word's toxicity score is derived by comparing its vector representation—in this case, obtained using FastText embeddings—against a reference "toxic" vector. Thus, words positioned farther to the right have higher cosine similarity to this reference, indicating greater alignment with toxic or hateful language.

The color coding of words reflects their thematic categories. For instance, political terms (green), economic terms (blue), and repair/maintenance-related terms (purple) cluster toward the left side of the figure. These are typically neutral words with low toxicity scores. On the other hand, known hateful or offensive words (marked in red) appear toward the right side, where toxicity scores are higher.

This clear spatial separation—neutral words to the left and hateful words to the right—suggests that the FastText-based toxicity estimation is effective. The embedding model successfully distinguishes between ordinary language and toxic language, placing more harmful or offensive terms closer to the reference toxic vector. This outcome supports the validity of the chosen methodology for identifying and quantifying the degree of toxicity in textual data.

4 Empirical Analysis

In this section, we conduct our main econometric analysis, leveraging two natural experiments on the platform: The first involves introducing a mutual rating system, and the second implements rating-based monetary incentives. These design changes, which make individual evaluation visible, are expected to influence the toxic behavior of users who are especially sensitive to social approval.

To establish causality, we employ a difference-in-differences (DiD) approach. We compare changes in behavior before and after each platform change across two groups: (1) users who clearly seek public approval, as indicated by their public social media profiles elsewhere, and (2) users without such profiles, who likely place less value on approval. By tracking how these two groups adjust their posting behavior once the rating system or monetary incentives are introduced, we isolate the impact of these design changes on the approval-seeking group.

4.1 Key Treatment Variables

Post-Treatment Dummies Our DiD approach employs two before-and-after indicators: one for introducing a mutual rating system and another for providing rating-based monetary incentives. For both treatments, we distinguish between ex-ante (pre-treatment) and ex-post (post-treatment) periods. Specifically, we create a variable that measures the number of days since the treatment's introduction. We do this by counting each 24-hour interval after the treatment's launch sequentially—assigning a value of 1 to the first 24 hours, 2 to the next 24 hours, and so forth. Before the treatment starts, the variable remains at 0 for all earlier days until 24 hours prior to the launch. Since our identification strategy focuses on a 30-day window (15 days before and 15 days after the treatment), this variable ultimately ranges from -15 to 15. We repeat this process for each treatment.

Using Public Profiles as a Proxy for the Strength of Need for Social Approval Our DiD approach incorporates an additional dimension by classifying users into a treatment group, characterized by a stronger tendency to seek public approval, and a control group, which likely places less emphasis on social validation. For this classification, we use the presence of public profiles on other social media as a proxy for the strength of the need for social approval. This approach is grounded in a well-established body of literature in social psychology. The following discussion elaborates on how the literature supports the use of public profiles as a reliable indicator of the strength of the need for social approval.

Public social media profiles signal users' desire for social validation, as they allow unrestricted visibility and enable users to seek validation from broader audiences through likes, comments, and shares (Tang, 2022; Meythaler et al., 2023; Sciara et al., 2023). By contrast, private profiles limit visibility and, hence, reduce the extent of public endorsement (Gruzd and Hernández-García, 2018). The underlying mechanisms resonate with core theories in social psychology, such as Social Comparison Theory (Festinger, 1957; Suls and Wheeler, 2012) and Uses and Gratifications Theory (Katz et al., 1973; Cipolletta et al., 2020). The theory tells us that when individuals strongly desire approval, they gravitate toward having public profiles that facilitate social comparisons, allowing them to gauge their self-worth through visible metrics.[5] In sum, identifying public profiles as a proxy for the strength of the need for social approval leverages this established research.

How to Detect Public Profiles Practically, we define "public profiles" as the presence of shared, publicly accessible SNS profiles. More precisely, our classification is based on either of two criteria: (i) whether the user's Q&A platform account page includes URL links to external social media (e.g., Twitter, Instagram, TikTok, Facebook), or (ii) whether the same username on the Q&A platform is used on other social media sites. Users are encouraged to register on our platform using an existing social media account (e.g., Twitter) since it allows users to avoid re-entering the required information during registration. Therefore, the same username frequently appears across different sites.[6]

The information to define the status of holding public profiles is retrieved one month before the first intervention. Importantly, their status of having (or not having) public profiles is highly stable: only five users opened new social media accounts with public profiles after either first or second natural experiments.

The Number of Users in Public Profiles and No Public Profiles Groups In total, 22,236 users posted a response at least once during 15 days before and after the

[5]Existing empirical findings further elucidate the relationship between public profiles and approval-seeking behavior. Valkenburg et al. (2006) demonstrate that positive validation via public profiles enhances feelings of satisfaction and belonging, while Jackson and Luchner (2018) find that negative feedback often prompts defensive behavior, including the shift to private profiles. Moreover, Burrow and Rainone (2017) show that the emotional responses tied to feedback are moderated by personal traits like the need for approval, magnifying the effects of social media interactions. They also find that high follower counts bolster self-esteem and motivate users to maintain their public disclosure habits, suggesting a cyclical reinforcement of public behavior. Li et al. (2018) indicate that users often employ public profiles to maximize visibility and engagement, a practice aligned with the Uses and Gratifications Theory.

[6]Note that we use "public profile" and "public account" interchangeably to describe an account accessible by anyone, regardless of whether it includes personal details. Even without disclosing private information, a user can still maintain a publicly accessible account.

first intervention. Figure A.2 shows how often 22,236 users posted during this period. Of these users, 13,208 (59.4%) have public profiles, and 9,028 (40.6%) do not. In the public-profile group, 1.28% posted five times, 2.84% posted four times, 5.92% posted three times, 12.38% posted twice, and 77.98% posted only once. In the no-public-profile group, 1.08% posted five times, 2.74% posted four times, 5.62% posted three times, 12.18% posted twice, and 77.78% posted only once.[7]

In total, 22,308 users posted a response at least once during 15 days before and after the second intervention. Figure A.3 displays posting frequencies for 22,308 users. Among these users, 13,260 (59.4%) have public profiles and 9,048 (40.6%) do not. In the public-profile group, 1.4% posted five times, 3.1% posted four times, 5.6% posted three times, 11.8% posted twice, and 78.1% posted only once. In the no-public-profile group, 1.2% posted five times, 2.9% posted four times, 5.4% posted three times, 12.2% posted twice, and 78.3% posted only once.[8]

4.2 Graphical Display of Causal Impacts

We hypothesize that users in the high social approval group—those who maintain publicly recognizable accounts—will exhibit stronger reactions to the introduction of the like-button and monetary incentives, as these users are especially sensitive to visible feedback and social validation.

Figure 1 presents two panels tracking changes in average toxicity scores among two distinct user groups before and after key platform design changes. On both graphs, the vertical dashed line at day zero marks the introduction of a new feature—first the like-button system and then rating-based monetary rewards. Time is measured in days, and toxicity is represented by an average cosine similarity metric, which captures how closely users' content matches a known set of toxic speech patterns. Note that we standardize this toxic score to make its mean value zero and make its standard deviation one. Before each intervention, the two groups—users with public social media profiles and those without—exhibit similar patterns, with toxicity scores hovering around neutral values. This indicates that prior to the treatment, both groups trended similarly, creating a valid baseline for a DiD approach. Immediately after the introduction of the like-button system in the first panel, a noticeable divergence emerges. Public profile users begin producing

[7]We use two-proportion z-tests to check whether posting frequencies differ between the public-profile and no-public-profile groups. We find no statistically significant differences between the two groups for each number of posts.

[8]We again employ two-proportion z-tests to assess differences in posting behavior between public-profile and no-public-profile users. We find no evidence of systematic differences between the two groups, even under the new monetary reward structure.

more toxic content, evidenced by their toxicity scores rising well above zero in the days following the change. In contrast, users without public profiles remain relatively stable, indicating that the new rating mechanism primarily affects those who care more about external validation.

A similar pattern appears in the second panel following the introduction of monetary incentives based on user ratings. Again, public profile users respond by substantially increasing their toxicity levels. Users without public profiles do not mirror this increase, further reinforcing the idea that the design changes are pushing an approval-sensitive subset of the user base toward more negative, inflammatory content.

4.3 A Difference-in-Differences Approach

4.3.1 Identification Strategy

Intuitive Explanation Our primary objective is to estimate how two natural experiments—the introduction of a mutual rating system and the subsequent implementation of rating-based monetary rewards—affect the toxicity of content posted by users who seek social approval. For this purpose, we compare changes in the outcomes of approval-seeking users (those with a public social media profile at the baseline) to changes in non-approval-seeking users (those without such profiles) before and after each platform intervention. Note that we classify users as approval-seekers if they had a public social media account one month *before* the introduction of the like-button system, eliminating the influence of treatments on the public account status.

Time Windows A key identification challenge is determining whether any observed increase in toxicity arises from respondents adjusting their behavior in response to the new design, rather than from questioners posing more provocative questions that elicit toxic responses. To address this issue, we exploit the platform's feature that each question remains open for 15 days. We focus on responses to questions posted before the system changes but still open afterward. This approach ensures that we compare responses posted in the same question threads before and after the new design is introduced. By holding the question threads constant, we can attribute any change in toxicity to shifts in respondents' behavior rather than to differences in the underlying questions.

Specification We estimate the following model:

$$\begin{aligned} \text{ToxicScore}_{a,i,q,d} &= \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i \\ &+ \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}, \end{aligned} \tag{1}$$

where, the subscripts a, i, q, and d refer to answer, respondent, question, and date of posting, respectively. The outcome $\text{ToxicScore}_{a,i,q,d}$ measures toxicity as the cosine similarity between a respondent's post and a set of previously deleted, ethically problematic posts. We standardize this variable such that its mean value is equal to zero and standard deviation is equal to one. PublicAccount_i is an indicator variable equal to 1 if user i maintains a public social media account (and is therefore hypothesized to be more sensitive to social approval). The variable After_d equals 1 if the response was posted after the new system features were introduced and 0 otherwise. The vector X_i includes respondent characteristics such as age and gender. We include question fixed effects (θ_q) to control for all time-invariant question-specific factors, ensuring that we identify the causality within the same question threads. We also include date fixed effects (ζ_d) to account for time-varying common shocks. Standard errors are clustered at the question thread level.

The parameter of interest is δ: the interaction term $\text{After}_d \times \text{PublicAccount}_i$ captures the key effect of interest. A positive and significant δ would indicate that respondents with a public account become more toxic following the introduction of the mutual rating system or the rating-based monetary rewards.

The key identifying assumption for this DiD approach is that, in the absence of the platform interventions, the toxicity trends for both the approval-seeking and non-approval-seeking groups would have continued along parallel paths. Under this assumption, any post-treatment divergence can be attributed to the social approval incentives introduced by the platform changes.

4.3.2 Main Results

In Table 1, columns (1) and (2) present the estimates for the first natural experiment, in which the platform introduces a like-button system that makes social feedback visible to all. Column (1) incorporates question fixed effects and date fixed effects, while column (2) adds hour of the day fixed effects. In both columns, the key interaction term, which captures how users with a public social media profile respond to the introduction of the like-button system, is positive and highly statistically significant. The magnitude, consistently around 0.22 standard deviations, indicates that making social approval visible leads approval-seeking users to produce more toxic responses relative to their counterparts

without a public profile.

Columns (3) and (4) present the results for the second natural experiment, where the platform ties monetary rewards to user ratings. Both columns include question fixed effects and other relevant controls, with column (4) adding hour-of-the-day fixed effects. The results show that after the introduction of rating-based monetary rewards, approval-seeking users increase their toxicity scores by roughly 0.49 standard deviations more than their non-approval-seeking counterparts. The larger coefficient compared to the first intervention indicates that tying financial rewards to user ratings accentuates the upward shift in toxic content among those who are sensitive to social validation.

Overall, when the platform redesigns its interface first by making social approval signals visible, and later by adding monetary rewards, those who were already sensitive to public visibility shift their behavior and produce content that is more akin to previously removed, ethically problematic posts.

4.3.3 The Event Study Results

We implement the event-study style analysis where we compare the difference of toxicity between public profile and no public profile users in each event time relative to the reference event time. Panels (a) and (b) in Figure 2) plot coefficients for each event time: (a) the introduction of a mutual rating ("like-button") system; (b) the introduction of rating-based monetary rewards. Both figures use a similar framework, where the x-axis marks time intervals (aggregated into two-day bins) relative to the policy intervention date. The vertical dashed line in each figure indicates the moment the new feature was introduced. Shaded areas represent 95% confidence intervals.

In both panels, the pre-treatment coefficients are close to zero and statistically indistinguishable from zero. This indicates that before the introduction of the like-button system (left panel) or the monetary rewards (right panel), users with public accounts and those without were exhibiting parallel trends in toxicity. These results are consistent with the parallel trend assumption.

The post-treatment coefficients reveal how the interventions affect toxicity. In the left panel, immediately after the introduction of the mutual rating system, there is a noticeable jump in toxicity among users with public accounts compared to no public account users. This increase is statistically significant and persists over time. In the right panel, the introduction of rating-based monetary rewards generates an even larger and more immediate spike in toxicity. Again, this effect endures throughout the observation period. These patterns suggest that users who value social approval become more toxic

in their posts when they can gain approval through likes or when higher ratings translate into financial benefits.

4.3.4 Robustness Checks

Different Bandwidths Since we focus on responses to questions posted 15 days before and after the treatments in our main estimation strategy, one might be concerned that our results are driven by arbitrary choices of bandwidths. To deal with this concern, we perform a robustness check to demonstrate that the estimation results are not bandwidth-dependent. Figure A.5 document the estimation results based on 4- to 15-day bandwidths for the first and second treatments, respectively. All coefficients are fairly robust and statistically significant with the same signs as the main results, and thus, it implies that our main results do not depend on the choice of bandwidths.

The Alternative Outcome: Postings Deleted by the Platform Here, we examine whether our estimation results are vulnerable to how the outcome is constructed. For this purpose, instead of the machine learning-based outcome, we use the dummy variable that takes the value of one if a response was forcefully deleted by the management platform due to an ethical violation. This dummy can capture whether a response contains inappropriate language toward other users. Table A.3 presents the estimation results based on Equation (1) but using the dummy variable for deleted postings by the management company as an outcome instead of the toxicity score. Estimation results show that the first treatment increases the probability of being deleted due to the rule violation by 0.99 to 0.97 percentage points, while the second treatment increases it by 1.52 percentage points. The scale of coefficients is economically prominent, given that the mean outcome value is 3.23% before the first treatment. Taken together, the system design changes induce users sensitive to social approval to use inappropriate language and thus increase the number of postings deleted by the management.

4.3.5 Treats to the Internal Validity

The Stable Unit Treatment Value Assumption (SUTVA) We must also examine whether the Stable Unit Treatment Value Assumption (SUTVA) holds. In our context, SUTVA would be violated if spillover effects exist between treated and control users: if users with public social media accounts (the treatment group) affect outcomes among users without public accounts (the control group), the control group would no longer provide a valid counterfactual, compromising the DiD framework's internal validity.

Construction of Two Variables for Measuring Exposure to Toxicity To test the validity of this assumption, we develop two measures of toxicity exposure. First, for each control user, we calculate the share of treated users in the 10-minute interval immediately before they post on a given thread. This measure captures the extent of exposure to treated users precisely when control users decide to post.

However, relying solely on the share of treated users is an indirect measure of exposure to toxicity for control users: merely logging in does not ensure any interaction between treated and control users. Therefore, we introduce a second measure: the share of toxic content in the same 10-minute window in a given question thread. This measure directly focus on control users' exposure to toxicity in their question threads.

Specification We focus on a sub-sample of responses posted exclusively by control users. We then estimate the following DiD model using one of these exposure measures:

$$\begin{aligned} \text{ToxicScore}_{a,i,q,d} &= \alpha + \beta\, \text{ToxicExposure}_{a,i,q,d} + \delta\, \text{After}_d \times \text{ToxicExposure}_{a,i,q,d} \\ &+ \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}, \end{aligned} \tag{2}$$

where $\text{ToxicExposure}_{a,i,q,d}$ represents either the share of toxic users or the share of toxic content in question q during the 10 minutes immediately preceding user i's post a on date d. All other variables are the same as those in Equation (1).

Estimation Results Table A.4 presents the estimation results. Columns (1) and (2) focus on a sub-sample of control users during the first intervention. In Column (1), the coefficients for both the share of treated users and its interaction with a post-dummy are near zero and statistically insignificant. Column (2) shows similarly negligible effects when using the share of toxic content in a thread as the exposure measure. These findings imply no evidence of spillovers before or after the first intervention: control users do not become more toxic when treated users are present more often or when toxic content in a thread is more prevalent. Consequently, control users appear unaffected by others' treatment status in this period.

Columns (3) and (4) present results for the introduction of the rating-based monetary reward and again point to no spillover effects. Control users' toxicity scores do not increase when they face a higher proportion of treated users or toxic content. Neither the baseline level of toxic exposure nor its interaction with the post-intervention period significantly raises toxicity among control users. These findings further support SUTVA in our setting, indicating that control users remain largely unresponsive to either the presence of treated

users or the volume of toxic content they produce.

4.3.6 Further Threats to Internal Validity

Ruling Out Post-Treatment Selection Here, we must verify whether the observed post-intervention differences in outcomes stem from actual behavioral shifts or from changes in the composition of either the treated or control groups. For instance, if certain types of users (distinguished by demographics or platform experience) disproportionately joined or left either group after the intervention, the resulting outcome differences may merely reflect who is in each group rather than actual behavioral changes by the same individuals.

First, we examine the stability of users' login behavior around each intervention period. Figure A.4 shows the proportion of logged-in users by treatment status during the 15-day window before and after each system change. We calculate this proportion by dividing the number of logged-in users on a given day by the total number of users observed over the entire period. The patterns indicate that the shares of both public-profile and non-public-profile users remain stable around the introduction of the like button and the rating-based monetary reward. In other words, there are no substantial shifts in login rates, suggesting that the interventions did not cause different types of users to join or leave disproportionately.

To further address post-treatment selection, we conduct balance tests on predetermined user characteristics using the full DiD specification. If the interventions had caused treated users to differ systematically from their control counterparts on these baseline traits, the coefficients on these interaction terms would be significant. However, as shown in Tables A.5 and A.6, these interaction terms remain statistically indistinguishable from zero. This finding confirms that treated users' characteristics relative to the control group do not shift after the interventions.

In sum, both the logged-in user proportions and the balance tests on user characteristics reveal no evidence of selective attrition or composition changes. These results suggest that the observed changes in toxicity among public-account users reflect genuine behavioral shifts in response to the platform interventions, rather than changes in which users comprise the treated group.

Possible Contamination by Ordering Effects Another potential concern for our identification strategy involves the sequence in which responses appear within the same thread. By construction, all posts classified as "post-treatment" are placed in threads where at least one "pre-treatment" or even another "post-treatment" post has already

been submitted. Under this scenario, one could argue that the observed increase in toxicity among public-profile users after the intervention might simply reflect an "ordering effect"—that is, the possibility that later responses in a thread tend to be more extreme regardless of treatment status.

To address this alternative explanation, we augment the baseline specification by including order fixed effects, controlling for the number of posts already present in a thread before a given response is submitted. This approach absorbs any observable or unobservable features related to the placement of a response within the sequence of existing posts, ensuring that what we attribute to the policy change is not merely a by-product of posting order.

Table A.7 presents the estimation results with these additional order controls. As shown, the coefficients for the interaction between having a public profile and the post-intervention period remain positive and statistically significant. In all cases, the inclusion of order fixed effects does not diminish the magnitude or significance of the key interaction terms. These findings indicate that the documented increase in toxicity for public-profile users following the platform interventions is robust to controlling for the sequence in which responses appear, ruling out the ordering effect as an alternative explanation.

4.3.7 Interpretation

Does Having a Public Account Really Indicate Social Approval Seeking? A central question arises from our findings: does the observed increase in toxicity among users identified as "public-account holders" truly stem from their heightened sensitivity to social approval, or could it simply reflect some intrinsic characteristic correlated with holding a public account? If having a public account—regardless of its level of activity—captures unobservable traits (e.g., personality type, baseline curiosity in public visibility) that are unrelated to social approval-seeking preferences, then attributing the toxicity increase to the social approval incentives could be misleading.

To test the validity of this concern, we distinguish between "active" and "inactive" public account holders. An inactive public account is a publicly accessible but lacks any actual posts or updates for the last 12 months. Such users, by definition, are far less likely to be actively seeking approval or engagement from an audience, as they do not produce any content. If the mere fact of having a public account drives the observed changes in toxicity, inactive public account holders should exhibit patterns similar to active ones after the introduction of new system designs.

However, the results in Table 2 consistently reject this alternative interpretation. The

coefficients for inactive public account holders are near zero and statistically insignificant, indicating no discernible change in their toxicity patterns after the platform interventions. These null results imply that behavioral changes observed in the main analysis are actually driven by the desire for public approval rather than by intrinsic user attributes associated with having a public account.

Non-anonymity Does Not Explain Our Results Another possible interpretation is that what we describe as "approval-seeking" behavior could instead reflect reputational concerns driven by non-anonymity. Specifically, having a public account might expose a user's identity, heightening their reputational concerns in online communities. Furthermore, new system designs that emphasize visibility and reputation metrics may exacerbate these concerns. Therefore, this alternative explanation suggests that if users are incentivized to align with social norms (e.g., average toxicity in this context), the observed increase in toxicity could stem not from a desire for social approval but from intrinsic motivations to comply with community norms to avoid reputational punishment. Consequently, it predicts that less anonymous users might adopt more aggressive or toxic behaviors to conform to updated community norms where toxic comments become more prevalent, particularly after the introduction of new system designs.

To investigate this alternative explanation, we construct an indicator to directly measure the degree of anonymity. Users who provide identifiable information, such as job titles, affiliations, or other private details, on their account pages—either on the Q&A platform or other social media—are classified as "non-anonymous." Importantly, having a public account (our main treatment) does not inherently imply non-anonymity. A public account on social media is defined as using the same username across platforms, but non-anonymity specifically requires the disclosure of private information in at least one online community. For instance, a user with a public account on social media may still choose not to disclose private information on that account. Conversely, a user without a public account on social media might disclose private affiliations on their Q&A platform account. If the observed treatment effects of having a public account were merely driven by non-anonymity, we would expect non-anonymous users to exhibit a similar increase in toxicity following the introduction of new system designs.

We define each user's non-anonymity status based on the information before the first intervention. At baseline, 12.34 percent of users disclose private information in at least one online community, classifying them as non-anonymous. Following the first intervention, we track changes in users' anonymity status and find that only three users shift their status by withdrawing their disclosed information. This negligible change in anonymity

status over time justifies our use of baseline non-anonymity as the treatment variable for subsequent analyses.

The results in Table 3 clearly reject this alternative explanation. The coefficients for non-anonymous users under post-treatment conditions are near zero and statistically insignificant, indicating that non-anonymity alone does not lead to the observed behavioral patterns. Specifically, while users with public profiles exhibit a significant increase in toxicity after the platform interventions, non-anonymous users show no such change throughout all specifications. These findings suggest that the key driver of increased toxicity is not simply the users' identifiability or reputational concerns stemming from disclosed personal information. Rather, the increase in toxicity appears to be tied to the heightened desire for social approval—a trait unique to users who actively attract audiences through public accounts.

Who Responds to System Design Changes? Potentially Harmful, but Previously Hidden Users Become More Toxic A natural follow-up question is: Which users are most responsive to the introduction of new system designs? We identify two mutually exclusive possibilities. First, some users may have a latent inclination toward toxic behavior yet have never violated the platform's rules. Before the new designs, these individuals remained within acceptable boundaries, making them indistinguishable from the wider user base (thus, they are potentially harmful but previously hidden). However, once driven by the new incentives and a heightened desire for social approval, they may now cross the line into toxicity. Second, users who have already committed rule violations may escalate further, adopting even more extreme or aggressive language.

We again use a machine learning approach, based on FastText word embeddings, to identify "potentially harmful" users. First, we create a reference subsample consisting of non-removed posts from users who had at least one post forcibly removed for toxic content in the pre-treatment period. Next, we measure the language similarity between these reference posts and each target post. Finally, we classify as "potentially harmful" those users whose posts have never been removed by the platform before the first intervention but whose average language similarity is above the median. Although these users are never toxic themselves, they display language patterns similar to those used by already-toxic users.

We split our sample into three sub-samples: (i) Answers posted by users whose posts had never been removed by the platform before the first intervention and are identified as "potentially harmful" by the ML algorithm, (ii) answers posted by users whose posts had never been removed by the platform before the first intervention and are not identified as

"potentially harmful" by the ML algorithm; and (iii) answers posted by users whose posts had been removed by the platform at least once before the first intervention.

We then run separate DiD estimations on each of the three subsamples to determine which group is most affected by the system design changes. As shown in Table 4, potentially harmful users exhibit a notable increase in toxicity once the new designs are introduced. In particular, users whose language patterns already resembled those of individuals producing extreme content become significantly more toxic when the new incentives take effect (Columns 1 and 4). In contrast, users who do not align with already-toxic individuals show no significant change under the same conditions (Columns 2 and 4). Finally, already-toxic users appear largely unaffected by the new system designs (Columns 3 and 6).

Using an additional approach, we examine whether already-toxic or never-toxic users respond more strongly to the interventions. We conduct a quantile regression with the baseline DiD specification. As shown in Figure 3, analyzing treatment effects at different points in the toxicity distribution reveals that less initially toxic users—those at lower quantiles—experience more pronounced increases in response to the new incentive structures. This finding aligns with the idea that users who were relatively moderate at baseline, rather than already highly toxic, are the ones adjusting their behavior most under the new system design. In other words, the like-button system and rating-based monetary rewards encourage these previously moderate yet potentially harmful users to escalate their language.

Taken together, these findings indicate that the new system design does not affect all users equally; rather, it primarily influences those who were already predisposed to more harmful behavior but remained hidden before the interventions.

4.3.8 Does Toxicity Really Attract More Attention?

Is the increased toxicity among social-approval-seeking users an effective way to attract a larger audience, or are they misperceiving the best strategy for gaining positive feedback? To investigate whether toxic behavior truly boosts popularity, we re-estimate our DiD model using the number of "likes" as the outcome variable. We categorize these likes based on their source—other toxic users, non-toxic users, questioners, and the overall user base—to evaluate how different audiences respond to newly toxic posts.

Table 5 presents the estimation results. After the introduction of the new system designs, public-profile users who become more toxic do indeed receive more likes, but only from other toxic users (Columns 1 and 5). Conversely, they earn fewer likes from non-toxic

users and questioners (Columns 2, 3 for the like-button system and 6, 7 for the monetary reward system), ultimately decreasing their total number of likes (Columns 4 and 8). In other words, although these approval-seeking individuals may think that increasing toxicity will improve their overall popularity, this strategy backfires when considering the broader community. Their posts alienate the majority of users and fail to generate a net increase in positive feedback. This finding suggests that these users "misperceive" the acceptable boundaries of platform behavior, mistakenly assuming that more extreme posts will lead to greater recognition.

4.4 Mechanisms: Misperceived Legitimization

So far, our analysis shows that once visible "likes" and rating-based monetary rewards are introduced, some public-profile users who are potentially harmful but originally hidden begin posting more toxic responses, despite that extreme behavior does not attract more attention in total.

Why do these users misunderstand the acceptable level of toxicity and become more toxic? We hypothesize that they revise their perceptions of social norms after seeing initially toxic content receive positive reactions from other users. On our Q&A platform, the top-page thread displays the most recent questions and responses, but it updates very frequently due to high posting activity. If a user visits and coincidentally sees a toxic comment with positive feedback, they may interpret this as a signal that such behavior is both permissible and desirable for attracting an audience. We call this phenomenon "misperceived legitimization," where the visibility and endorsement of toxic posts lead otherwise hidden users to believe that harsher language, which they might have previously avoided, is actually approved by the community.

We test this mechanism by examining each user's login time and the top-page comments displayed at that moment. Specifically, we identify which posts appear when users visit the platform and categorize their subsequent responses based on the type of comments listed at the top of the page. We then re-estimate our DiD framework using the same control group as before (i.e., posts published prior to the interventions) but define three categories of posts for the treatment period:

1. Posts from users who observed **no toxic comments** at the top of the page.
2. Posts from users who observed **toxic comments without any likes** at the top of the page.
3. Posts from users who observed **toxic comments with likes** at the top of the page.

By comparing these three groups, we determine whether seeing a liked toxic comment increases toxicity among approval-seeking users.

Table 6 displays our estimation results. Columns (1) and (4) show responses posted when no toxic posts appear on the top page; Columns (2) and (5) capture responses posted when toxic posts without likes are on the top page; finally, Columns (3) and (6) involve responses posted when toxic posts with likes are shown.

Our findings indicate that toxicity rises only when users see a "liked" toxic post on the top page. Specifically, when potentially harmful, approval-seeking users encounter a toxic comment that has already received likes, they significantly increase their own toxicity (see Columns 3 and 6). In contrast, encountering a non-toxic top-page comment or a toxic comment without likes does not yield a comparable increase in toxic behavior. These results imply that it is not merely the presence of toxic content but its positive endorsement ("liked" status) that leads hidden users to believe their own toxic behavior is legitimized.

This evidence supports the misperceived legitimization hypothesis. Potentially harmful users, previously unsure about acceptable boundaries, interpret "liked" toxic comments as endorsements of their latent negative tendencies. The platform's design changes—making likes visible and tying them to monetary rewards—intensify this misunderstanding. Consequently, these users mistake a vocal minority's positive reaction for broad community approval, prompting them to adopt increasingly toxic behavior over time.

5 The Formation of Segregated Communities

In this final section, we examine whether the platform's design interventions lead to the formation of segregated communities, or "echo chambers," characterized by similar levels of toxicity.

The Social Network Approach Based on The Follow Feature We leverage the "follow" feature, a common design element on many Q&A platforms and social media sites. On our target Q&A platform, following allows users to subscribe to another user's content—questions, answers, or discussions—or to particular topics, without requiring mutual consent. This setup helps individuals track posts they find relevant or interesting, resulting in a personalized content stream.

Technically, clicking a "Follow" button on a user's profile (or on a specific topic) prompts the system to record the follower–followed relationship in its database. Whenever the followed user posts new answers or updates existing ones, the follower receives

notifications or sees these updates in their newsfeed. Therefore, these follow relationships produce an observable, one-way network of social ties. Note that this follower-followed relationship is termed "friendships" in our platform. However, no mutual approval is necessary—users can add others as friends without their permission.

We use this feature to investigate whether, after the introduction of new platform system designs, users with similar toxicity levels become more likely to follow (or "befriend") each other. If so, it suggests that segregated clusters of like-minded users—potential echo chambers—may form.

Data Construction We construct a dyadic dataset, where each pair of users (i, j) constitutes one observation per month. For each pair, we create a friendship dummy variable taking the value of one if user j is listed on the friend list of user i in month m; we also create the absolute difference of toxicity between user i and j. Our analysis then examines how differences in toxicity between two users relate to the friendships before and after the platform's design changes.

Empirical Specification To investigate whether friendships increasingly involve users with similar toxicity levels, we estimate the following simple regression model:

$$\begin{aligned} \text{ToxicityDiff}_{i,j,m} \;=\; \alpha \;+\; \beta_1\, \text{Friendship}_{i,j,m} \;+\; \beta_2\, \text{Friendship}_{i,j,m} \times \text{PostChange}_m \\ +\, \gamma\, \mathbf{X}_{i,j,m} \;+\; \tau_m \;+\; u_{i,j,m}, \end{aligned} \tag{3}$$

where $\text{ToxicityDiff}_{i,j,m}$ is the absolute difference in toxicity scores between users i and j in month m, $\text{Friendship}_{i,j,m}$ indicates newly established friendships, PostChange_m denotes observations following the platform updates, $\mathbf{X}_{i,j,m}$ is a vector of dyad-level controls (e.g., shared content categories), τ_m represents month fixed effects, and $u_{i,j,m}$ is the error term. We also include user-level demographic proxies as part of $\mathbf{X}_{i,j,m}$.

Our coefficient of interest is the interaction term, β_2, which reflects whether friendship ties are more likely to involve users with similar toxicity scores after the platform changes. If this coefficient is negative, the results would imply that, once new system designs are in place, befriended users tend to have similar toxicity scores.

Estimation Results Table 7 presents the regression results, which support our hypothesis. First, before the initial intervention, users not in friendships do not differ significantly in toxicity from those who are in friendships (Column 1). Second, after the like-button

system is introduced, users in friendships display greater similarity in toxicity (Column 1). Third, following the implementation of monetary incentives, the degree of similarity in toxicity becomes even higher (Column 2). In other words, users increasingly form ties with friends who resemble them in terms of toxicity. This pattern implies that the new platform schemes not only raise toxicity levels for some users but also lead to the emergence of like-minded, more segregated communities based on toxicity.

We emphasize that this analysis reflects correlation rather than causation. Nevertheless, the post-change friendships appear to involve users exhibiting more similar toxicity, suggesting a shift toward more homogeneous and segregated communities once the new system designs are introduced.

6 Conclusion

This study demonstrates how platform designs emphasizing social approval can escalate online toxicity. By implementing mutual rating systems (e.g., like-buttons), platforms increase the visibility of approval metrics, prompting users with strong social approval-seeking tendencies to adopt more toxic behaviors. The subsequent introduction of monetary incentives further exacerbates this toxicity, as users amplify their harmful content in response to pecuniary rewards. We identify misperceived legitimization as the primary mechanism, in which users misinterpret visible endorsements of toxic content as broad societal approval, leading them to escalate their toxicity. This study offers a previously under-explored perspective on how platform design, monetary incentives, and non-monetary incentives intersect to exacerbate online toxicity.

Nevertheless, our findings should be interpreted with caution. Although the target platform shares many structural features with other major platforms—such as like buttons, user incentives, and similar communication frameworks—differences in user demographics, cultural context, and moderation policies may limit the generalizability of our conclusions. Future research should replicate these analyses on various platforms and in diverse cultural settings to establish a more comprehensive understanding of how design elements that foster social approval-seeking can drive toxicity.

References

Luis Abreu and Doh-Shin Jeon. Homophily in social media and news polarization. *NET Institute Working Paper*, 19-05, 2019.

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. A model of online misinformation. *Review of Economic Studies*, 2023. 10.1093/restud/rdad111.

Daron Acemoglu, Asuman Ozdaglar, and James Siderius. A model of online misinformation. *Review of Economic Studies*, 91(6):3117–3150, 2024.

Hunt Allcott, Luca Braghieri, Sarah Eichmeyer, and Matthew Gentzkow. The welfare effects of social media. *American Economic Review*, 110(3):629–76, 2020a.

Hunt Allcott, Matthew Gentzkow, and Chuan Yu. Polarization and public misinformation in the social media age. *Journal of Economic Perspectives*, 34(3):50–70, 2020b.

Guy Aridor, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. The economics of social media. *Journal of Economic Literature*, 62(4):1422–74, December 2024. 10.1257/jel.20241743. URL https://www.aeaweb.org/articles?id=10.1257/jel.20241743.

Solomon E Asch. Effects of group pressure upon the modification and distortion of judgments. In Harold Guetzkow, editor, *Groups, Leadership, and Men: Research in Human Relations*, pages 177–190. Carnegie Press, 1951. URL https://psycnet.apa.org/record/1952-00803-001.

Nava Ashraf and Oriana Bandiera. Social incentives in organizations. *Annual Review of Economics*, 10(1):439–463, 2018.

Christopher A. Bail, Luke P. Argyle, and Taylor W. Brown. Social media and political polarization in the united states. *Proceedings of the National Academy of Sciences*, 118 (50):e2115670118, 2021.

Roy F Baumeister and Mark R Leary. The need to belong: Desire for interpersonal attachments as a fundamental human motivation. *Psychological Bulletin*, 117(3):497, 1995. URL https://doi.org/10.1037/0033-2909.117.3.497.

Timothy Besley and Maitreesh Ghatak. Prosocial motivation and incentives. *Annual Review of Economics*, 10(1):411–438, 2018.

J Aislinn Bohren, Alex Imas, and Michael Rosenberg. The dynamics of discrimination: Theory and evidence. *American economic review*, 109(10):3395–3436, 2019.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146, 2017.

Pedro Bordalo, Katherine Coffman, Nicola Gennaioli, and Andrei Shleifer. Stereotypes. *Quarterly Journal of Economics*, 131(4):1753–1794, 2016.

Samuel Bowles. Policies designed for self-interested citizens may undermine" the moral sentiments": Evidence from economic experiments. *science*, 320(5883):1605–1609, 2008.

Samuel Bowles and Sandra Polania-Reyes. Economic incentives and social preferences: substitutes or complements? *Journal of economic literature*, 50(2):368–425, 2012.

Levi Boxell, Matthew Gentzkow, and Jesse M Shapiro. Greater internet use is not associated with faster growth in political polarization among us demographic groups. *Proceedings of the National Academy of Sciences*, 114(40):10612–10617, 2017.

Anthony L Burrow and Nicolette Rainone. How many likes did i get?: Purpose moderates links between positive social media feedback and self-esteem. *Journal of Experimental Social Psychology*, 69:232–236, 2017.

Leonardo Bursztyn and David Y. Yang. Misperceptions about others. *Annual Review of Economics*, 14:425–452, 2022.

Leonardo Bursztyn, Georgy Egorov, Ruben Enikolopov, and Maria Petrova. Social media and xenophobia: evidence from russia. *NBER WP*, 2019.

Leonardo Bursztyn, Alessandra L González, and David Yanagizawa-Drott. Misperceived social norms: Women working outside the home in saudi arabia. *American economic review*, 110(10):2997–3029, 2020.

Leonardo Bursztyn, Georgy Egorov, Ingar Haaland, Aakaash Rao, and Christopher Roth. Justifying dissent. *Quarterly Journal of Economics*, 138(3):1403–1451, 2023.

Gordon Burtch, Qinglai He, Yili Hong, and Dokyun Lee. How do peer awards motivate creative content? experimental evidence from reddit. *Management Science*, 68(5): 3488–3506, 2022.

Julia Cagé, Nicolas Hervé, and Béatrice Mazoyer. Social media influence mainstream media: Evidence from two billion tweets. *Sciences Po Research Paper*, 2022. http://sciencespo.hal.science/hal-03877907.

F. Campante, R. Durante, and A. Tesei. *The Political Economy of Social Media*. CEPR Press, Paris & London, 2023. URL https://cepr.org/publications/books-and-reports/political-economy-social-media.

Davide Cantoni, David Y. Yang, Noam Yuchtman, and Y. Jane Zhang. Protests as strategic games: Experimental evidence from hong kong's antiauthoritarian movement. *Quarterly Journal of Economics*, 134(2):1021–1077, 2019.

Yan Chen, F Maxwell Harper, Joseph Konstan, and Sherry Xin Li. Social comparisons and contributions to online communities: A field experiment on movielens. *American Economic Review*, 100(4):1358–98, 2010.

Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

Sabrina Cipolletta, Clelia Malighetti, Chiara Cenedese, and Andrea Spoto. How can adolescents benefit from the use of social networks? the igeneration on instagram. *International journal of environmental research and public health*, 17(19):6952, 2020.

Varad Deolankar, Jessica Fong, and S. Sriram. Content generation on social media: The role of negative peer feedback. *SSRN Working Paper*, 2023. http://dx.doi.org/10.2139/ssrn.4522092.

Paul DiMaggio, Eszter Hargittai, W Russell Neuman, and John P Robinson. Social implications of the internet. *Annual review of sociology*, 27(1):307–336, 2001.

Dean Eckles, René F. Kizilcec, and Eytan Bakshy. Estimating peer effects in networks with peer encouragement designs. *Proceedings of the National Academy of Sciences*, 113(27):7316–7322, 2016.

Florian Ederer, Paul Goldsmith-Pinkham, and Kyle Jensen. Anonymity and identity online, September 2024. URL https://florianederer.github.io/ejmr.pdf. Unpublished mimeo.

Lena Abou El-Komboz, Anna Kerkhof, and Johannes Loh. Platform partnership programs and content supply: Evidence from the youtube 'adpocalypse'. *CESifo Working Paper*, 2023.

Leon Festinger. A theory of social comparison processes. *Human Relations*, 7(2):117–140, 1954. URL https://doi.org/10.1177/001872675400700202.

Leon Festinger. Social comparison theory. *Selective Exposure Theory*, 16(401):3, 1957.

Apostolos Filippas, John Horton, and Elliot Lipnowski. The production and consumption of social media. In *Proceedings of the 23rd ACM Conference on Economics and Computation*, pages 287–288, 2022.

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

Uri Gneezy, Kenneth L Leonard, and John A List. Gender differences in competition: Evidence from a matrilineal and a patriarchal society. *Econometrica*, 77(5):1637–1664, 2009.

Sandra González-Bailón and Yphtach Lelkes. Do social media undermine social cohesion? a critical review. *Social Issues and Policy Review*, 17(1):155–180, 2023.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*. European Language Resources Association (ELRA), 2018. URL https://arxiv.org/abs/1802.06893.

Anatoliy Gruzd and Ángel Hernández-García. Privacy concerns and self-disclosure in private and public uses of social media. *Cyberpsychology, Behavior, and Social Networking*, 21(7):418–428, 2018.

Andrew Guess, Brendan Nyhan, and Jason Reifler. Echo chambers and polarization: Understanding their limits and effects. *Annual Review of Political Science*, 26:211–235, 2023a.

Andrew M Guess, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, and Matthew Gentzkow. How do social media feed algorithms affect attitudes and behavior in an election campaign? *Science*, 381(6656):398–404, 2023b.

Sergei Guriev, Emeric Henry, Théo Marquis, and Ekaterina Zhuravskaya. Curtailing false news, amplifying truth. *SSRN Working Paper*, 2023. 10.2139/ssrn.4616553.

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

Blake Hallinan and Jed R Brubaker. Living with everyday evaluations on social media platforms. *International Journal of Communication*, 15:19, 2021.

Yong Huang and Rajagopal Narayanan. The dynamics of engagement and platform effects in social networks. *Management Science*, 66(11):5192–5206, 2020.

Christina A Jackson and Andrew F Luchner. Self-presentation mediates the relationship between self-criticism and emotional response to instagram feedback. *Personality and Individual Differences*, 133:1–6, 2018.

Julie Jiang, Luca Luceri, Joseph B. Walther, and Emilio Ferrara. Social approval and network homophily as motivators of online toxicity. *arXiv preprint arXiv:2310.07779*, 2023. URL https://arxiv.org/abs/2310.07779.

Armand Joulin, Edouard Grave, Piotr Bojanowski, and Tomas Mikolov. Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 427–431, Valencia, Spain, 2017. Association for Computational Linguistics. URL https://arxiv.org/abs/1607.01759.

Elihu Katz, Jay G Blumler, and Michael Gurevitch. Uses and gratifications research. *The public opinion quarterly*, 37(4):509–523, 1973.

Anna Kerkhof. The impact of youtube's revenue-sharing programs on content creation. *Unpublished Manuscript*, 2024.

Timur Kuran. *Private truths, public lies: The social consequences of preference falsification*. Harvard University Press, 1997.

Yphtach Lelkes, Gaurav Sood, and Shanto Iyengar. The hostile audience: The effect of access to broadband internet on partisan affect. *American Journal of Political Science*, 61(1):5–20, 2017.

Tin Cheuk Leung and Koleman S. Strumpf. All the headlines that are fit to change: Analysis of headline changes in the media industry. *SSRN Working Paper*, 2023. http://dx.doi.org/10.2139/ssrn.4075398.

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

Pengxiang Li, Leanne Chang, Trudy Hui Hui Chua, and Renae Sze Ming Loh. "likes" as kpi: An examination of teenage girls' perspective on peer feedback on instagram and its influence on coping response. *Telematics and Informatics*, 35(7):1994–2005, 2018.

X. Ma. Incentives for digital content contribution. In *Social Influence on Digital Content Contribution and Consumption*, Management for Professionals. Springer, Singapore, 2023. 10.1007/978-981-99-6737-7_2. URL https://doi.org/10.1007/978-981-99-6737-7_2.

Antonia Meythaler, Hannes-Vincent Krause, Annika Baumann, Hanna Krasnova, and Jason Bennett Thatcher. The rise of metric-based digital status: an empirical investigation into the role of status perceptions in envy on social networking sites. *European Journal of Information Systems*, pages 1–28, 2023.

Roberto Mosquera, Mofioluwasademi Odunowo, Trent McNamara, Xiongfei Guo, and Ragan Petrie. The economic effects of facebook. *Experimental Economics*, 23:575–602, 2020.

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

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

Simha Mummalaneni, Hema Yoganarasimhan, and Varad Pathak. How do content producers respond to engagement on social media platforms? *SSRN Working Paper*, 2023. 10.2139/ssrn.4173537.

Ting Nie, Yanli Gui, and Yiying Huang. Online sharing behaviors driven by need for approval: the choice of individuals with low social intelligence and high gratitude? *Humanities and Social Sciences Communications*, 11:18, 2024. 10.1057/s41599-023-02535-8. URL https://doi.org/10.1057/s41599-023-02535-8.

Eli Pariser. *The filter bubble: How the new personalized web is changing what we read and how we think*. Penguin, 2011.

Markus Prior. Media and political polarization. *Annual Review of Political Science*, 16: 101–127, 2013.

Simona Sciara, Federico Contu, Mariavittoria Bianchini, Marta Chiocchi, and Giacomo Giorgio Sonnewald. Going public on social media: The effects of thousands of instagram followers on users with a high need for social approval. *Current Psychology*, 42:8206–8220, 2023.

Karthik Srinivasan. Paying attention. *Unpublished Manuscript*, 2023. https://karthikecon.github.io/karthiksrinivasan.org/paying_attention.pdf.

Jerry Suls and Ladd Wheeler. Social comparison theory. *Handbook of theories of social psychology*, 1:460–482, 2012.

Cass R Sunstein. *# Republic: Divided democracy in the age of social media*. Princeton University Press, 2018.

Jack Lipei Tang. Are you getting likes as anticipated? untangling the relationship between received likes, social support from friends, and mental health via expectancy violation theory. *Journal of Broadcasting & Electronic Media*, 66(2):340–360, 2022.

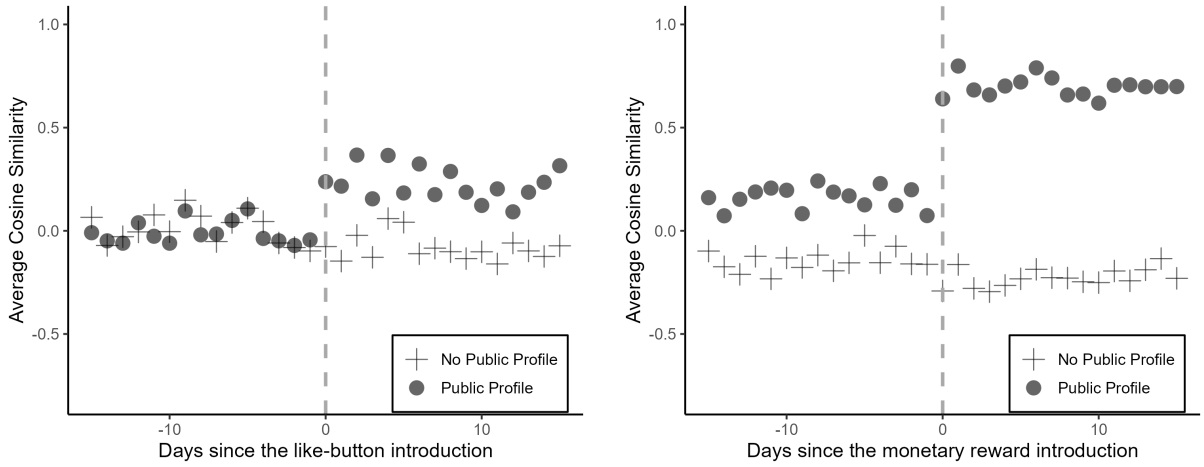
Patti M Valkenburg, Jochen Peter, and Alexander P Schouten. Friend networking sites and their relationship to adolescents' well-being and social self-esteem. *CyberPsychology & behavior*, 9(5):584–590, 2006.

Joseph B Walther. Social media and online hate. *Current Opinion in Psychology*, 45: 101298, 2022.

Zhiyu Zeng, Hengchen Dai, Dennis J. Zhang, et al. The impact of social nudges on user-generated content for social network platforms. *Management Science*, 69(9):5189–5208, 2023.

Ekaterina Zhuravskaya, Maria Petrova, and Ruben Enikolopov. Political effects of the internet and social media. *Annual Review of Economics*, 12:415–438, 2020.

Figure 1: The Toxicity Score Before and After the System Changes: the Introduction of the Like-button System and Rating-based Monetary Reward

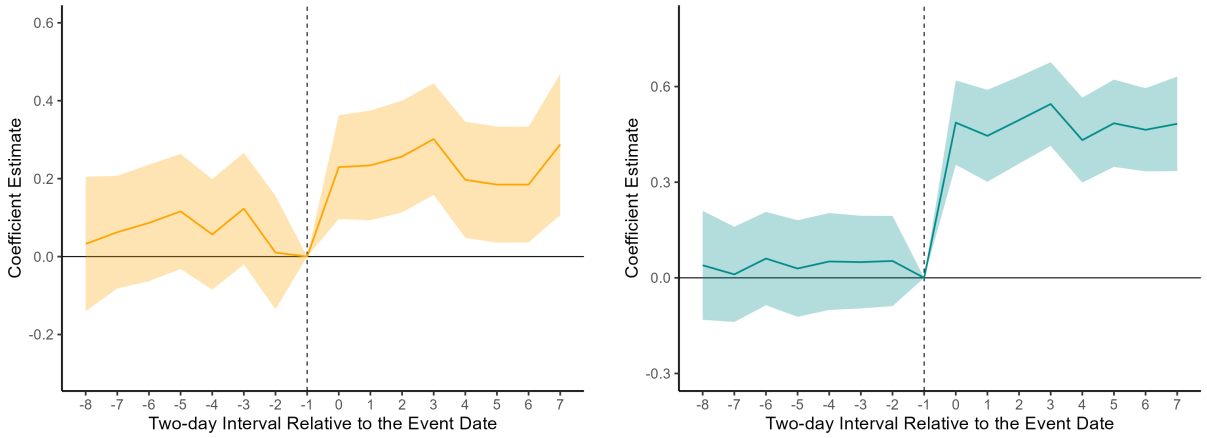


(a) The Introduction of the Like-button

(b) The Introduction of Monetary Rewards

Note: These figures show changes in the toxicity score over time, focusing on the 15 days before and after each system change treatment. The circle dots represent the average toxicity score for users with a public SNS profile outside the platform, while the plus marks represent the scores for users without such profiles. The x-axis indicates the number of days relative to the treatment date.

Figure 2: Event Study: The Impacts of the Introduction of the Like-button System and Monetary Reward

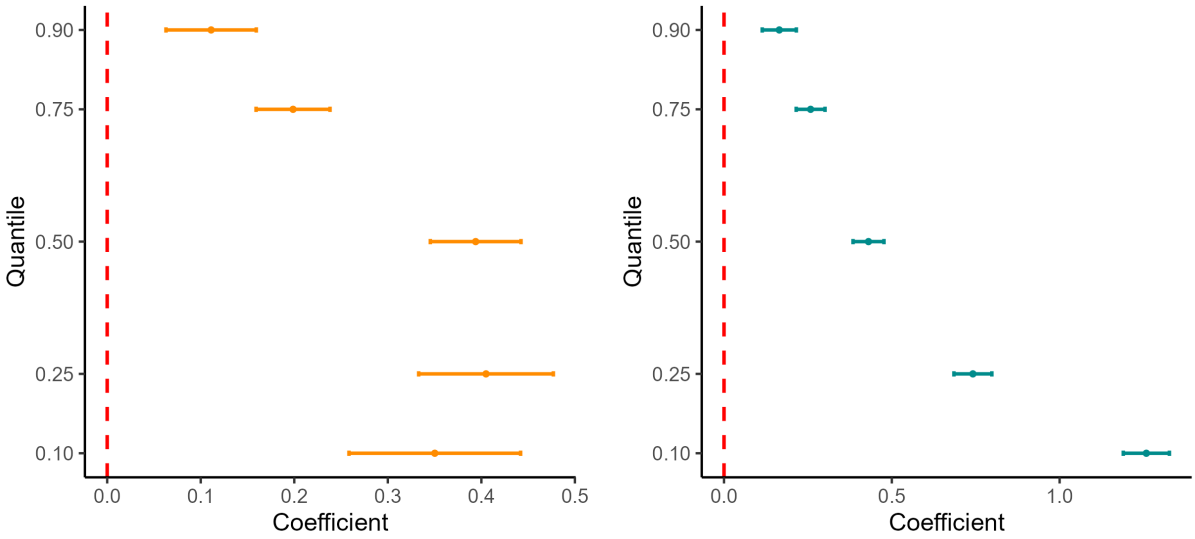


(a) The Introduction of the Like-button

(b) The Introduction of Monetary Rewards

Note: This figure presents the coefficients and 95% CI based on event study of the DiD approach: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\text{PublicAccount}_i + \sum_{d-j} \pi_j D_{d-j} \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,t,d}$ where PublicAccount indicates the dummy variable taking the value of one if user i has a public account profile in SNS outside the platform and zero otherwise. D_{d-j} is an indicator variable for event time j, meaning that the event took place j periods before this observation's calendar date. Note that, regarding event time, we use the two-day interval relative to the event date. The reference point is one or two days right before the treatment timing, which corresponds to -1 on the x-axis. We include question fixed effects and date fixed effects to control for question-specific factors and time-variant common factors.

Figure 3: Quantile Regression: The Distributional Impacts of the Introduction of the Like-button System and Monetary Reward



(a) The Introduction of the Like-button

(b) The Introduction of Monetary Rewards

Note: This figure presents the coefficients and 95% confidence intervals (CI) for the quantile regression models estimated at various quantiles ($\tau = 0.1, 0.25, 0.5, 0.75, 0.9$). The quantile regression models are specified as: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable captures the cosine similarity of user responses to removed answers, and the regressions control for text length. The dataset includes only observations within 15-dat windows before and after the natural experiments.

Table 1: The Impact of Introducing the Like-button System and Rating-based Monetary Rewards

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform			
	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-like-button Dummy	0.2215***	0.2214***		
	(0.0205)	(0.0205)		
Having a Public Profile × Post-monetary-reward Dummy			0.4881***	0.4880***
			(0.0221)	(0.0222)
R^2	0.28662	0.28704	0.32135	0.32200
Observations	30,485	30,485	30,640	30,640
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects		✓		✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score is estimated by a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1) and (2) include all answers posted within 15 day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted within 15 ay windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 2: Placebo Tests Using Inactive Public Profiles in SNS Outside of the Platform: The Impact of Introducing the Like-button System and Rating-based Monetary Rewards

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform			
	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	(1)	(2)	(3)	(4)
Having a Inactive Public Profile × Post-like-button Dummy	-0.0502	-0.0516		
	(0.0418)	(0.0419)		
Having a Inactive Public Profile × Post-monetary-reward Dummy			-0.0031	-0.0012
			(0.0435)	(0.0436)
R^2	0.28104	0.28146	0.27137	0.27209
Observations	30,485	30,485	30,640	30,640
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects		✓		✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\text{PublicInactiveAccount}_i + \delta\text{After}_d \times \text{PublicInactiveAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score, estimated using a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1) and (2) contain all answers posted within two-week windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted within two-week windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 3: Alternative Explanations: Non-anonymity Does Not Explain Increased Toxicity After the System Design Changes

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform			
	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-like-button Dummy	0.2214***	0.2213***		
	(0.0205)	(0.0205)		
Non-anonymous User × Post-like-button Dummy	-0.0017	-0.0003		
	(0.0484)	(0.0483)		
Having a Public Profile × Post-monetary-reward Dummy			0.4877***	0.4876***
			(0.0221)	(0.0222)
Non-anonymous User × Post-monetary-reward Dummy			0.0795	0.0788
			(0.0527)	(0.0527)
R^2	0.28666	0.28708	0.32144	0.32209
Observations	30,485	30,485	30,640	30,640
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects		✓		✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta_1 \text{PublicAccount}_i + \delta_1 \text{After}_d \times \text{PublicAccount}_i + \beta_2 \text{NonAnonymous}_i + \delta_1 \text{After}_d \times \text{NonAnonymous}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score is estimated by a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1) and (2) contain all answers posted two weeks before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted two weeks before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 4: Interpretation: Potentially Harmful Users Detected by the Machine Learning Model Respond to System Changes

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform					
	Introduction of Like-button System			Introduction of Rating-based Monetary Reward		
	Users Detected As Potentially Harmful	Users Not Detected As Potentially Harmful	Users Who Previously Violated a Rule	Users Detected As Potentially Harmful	Users Not Detected As Potentially Harmful	Users Who Previously Violated a Rule
	(1)	(2)	(3)	(4)	(5)	(6)
Having a Public Profile × Post-like-button Dummy	0.1059***	0.0108	0.0524			
	(0.0334)	(0.0242)	(0.0555)			
Having a Public Profile × Post-monetary-reward Dummy				1.283***	-0.0065	0.0029
				(0.0577)	(0.0264)	(0.0686)
R^2	0.41780	0.49517	0.67452	0.46696	0.42887	0.63871
Observations	13,565	13,011	3,908	13,634	13,550	3,455
Question fixed effects	✓	✓	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓	✓	✓
Hour of the Day fixed effects		✓	✓		✓	✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\text{PublicAccount}_i + \delta\text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1), (2), and (3) include answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (4), (5), and (6) include answers posted within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors, clustered at the question thread ID level, are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table 5: Does Toxicity Attract More Attention? The Impact of Posting Toxicity on Likes Per Post

	Introduction of Like-button System				Introduction of Rating-based Monetary Reward			
	# of Likes From Toxic Users	# of Likes From Non Toxic Users	# of Likes From Questioners	# of Likes From Total Users	# of Likes From Toxic Users	# of Likes From Non Toxic Users	# of Likes From Questioners	# of Likes From Total Users
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Having a Public Profile × Post-like-button Dummy	0.0783***	-0.0224***	-0.0236***	-0.0174***				
	(0.0074)	(0.0039)	(0.0042)	(0.0037)				
Having a Public Profile × Post-monetary-reward Dummy					0.1596***	-0.0488***	-0.0535***	-0.0384***
					(0.0122)	(0.0054)	(0.0057)	(0.0051)
R^2	0.38383	0.15200	0.16373	0.16030	0.26952	0.13283	0.13958	0.12251
Observations	30,485	30,485	30,485	30,485	30,640	30,640	30,640	30,640
Question fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓	✓	✓	✓	✓
Hour of the Day fixed effects	✓	✓	✓	✓	✓	✓	✓	✓

Notes: This table presents regression results based on the following equation: $extLikes_{a,i,q,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the number of likes expressed by other users. Columns (1) and (2) contain all answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table 6: Mechanisms: Misperceived Legitimization Induces Users to Post Toxic Comments

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform					
	Introduction of Like-button System			Introduction of Rating-based Monetary Reward		
	No Toxic Post on Top Page	Non-Liked Toxic Post on Top Page	Liked Toxic Post on Top Page	No Toxic Post on Top Page	Non-Liked Toxic Post on Top Page	Liked Toxic Post on Top Page
	(1)	(2)	(3)	(4)	(5)	(6)
Having a Public Profile × Post-like-button Dummy	0.0034	-0.0063	0.2303***			
	(0.0325)	(0.0315)	(0.0231)			
Having a Public Profile × Post-monetary-reward Dummy				-0.0042	-0.0014	0.7289***
				(0.0333)	(0.0328)	(0.0261)
R^2	0.44395	0.44414	0.29360	0.41657	0.41791	0.34340
Observations	16,766	16,757	26,740	17,579	17,454	25,863
Question fixed effects	✓	✓	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓	✓	✓
Hour of the Day fixed effects			✓			✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\text{PublicAccount}_i + \delta\text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score is estimated by a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Column (1) includes answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). However, answers posted after the treatment event are restricted to those posted when no toxic answers by others are on the top page. In Column (2), answers posted after the treatment event are restricted to those posted when toxic answers NOT liked by others are on the top page. In Column (3), answers posted after the treatment event are restricted to those posted when toxic answers liked by others are on the top page. In Column (4), answers posted within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). However, answers posted after the treatment event are restricted to those posted when no toxic answers by others are on the top page. In Column (5), answers posted after the treatment event are restricted to those posted when toxic answers NOT liked by others are on the top page. In Column (6), answers posted after the treatment event are restricted to those posted when toxic answers liked by others are on the top page. In all specifications, the number of characters in a post is included as a control variable. Standard errors, clustered at the question thread ID level, are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

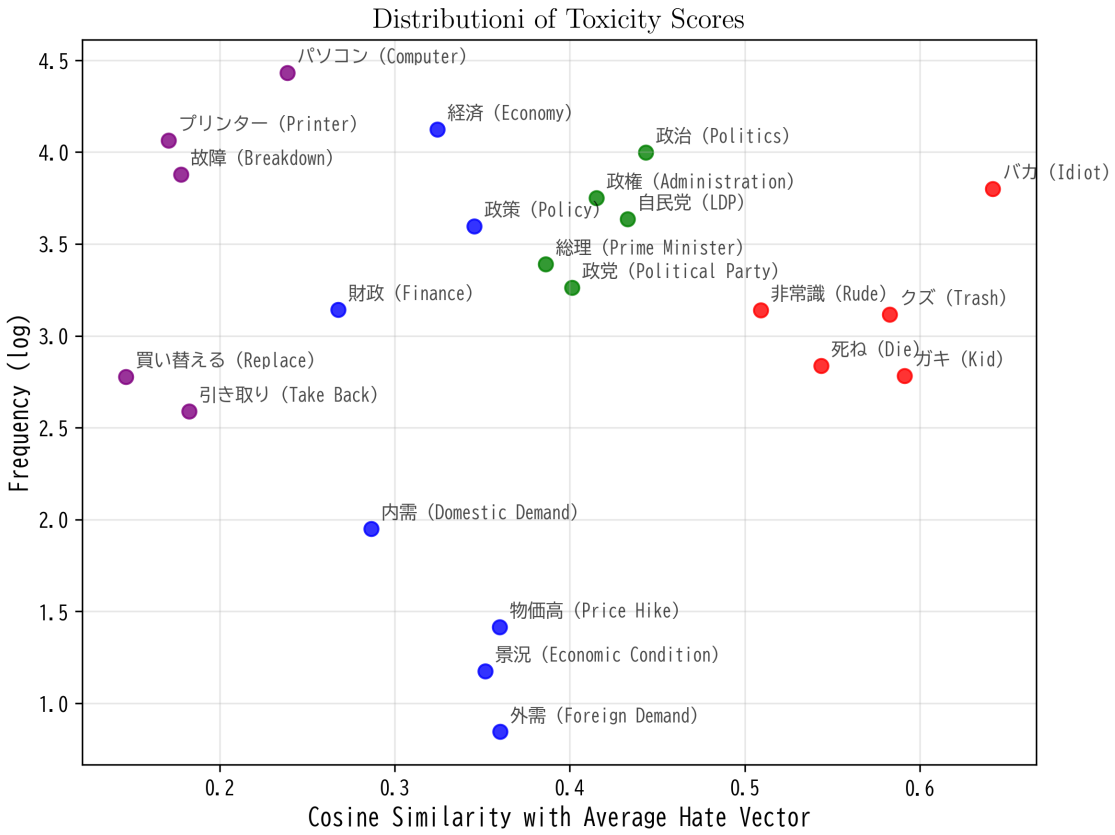
Table 7: Formation of Echo Chambers: The Impact of System Changes on Similarity of Toxicity Between Users Who Are Registered as Friends

	Dep Var: Similarity of Toxicity Between Users	
	Introduction of Like-button System	Introduction of Rating-based Monetary Reward
	(1)	(2)
Friend Registration Dummy	0.0017	-0.0285***
	(0.0016)	(0.0016)
Friend Registration Dummy × Post-like-button Dummy	-0.0303***	
	(0.0022)	
Friend Registration Dummy × Post-monetary-reward Dummy		-0.0296***
		(0.0022)
R^2	0.00023	0.00083
Observations	2,511,040	2,511,040

Notes: This table presents regression results based on the following equation: $similarity_{i,j} = \alpha + \beta friend_{i,j} + \delta after_{i,j} + \gamma friend_{i,j} \times after_{i,j} + \varepsilon_{i,j}$. The dependent variable is the similarity of average toxic scores between two users. The friend dummy variable is a binary indicator of whether user i registers user j as a friend in the platform. The after variable is a dummy taking the value of one for a pair of users after the introduction of the like-button system or the rating-based monetary incentives. In all specifications, standard errors clustered at a user ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

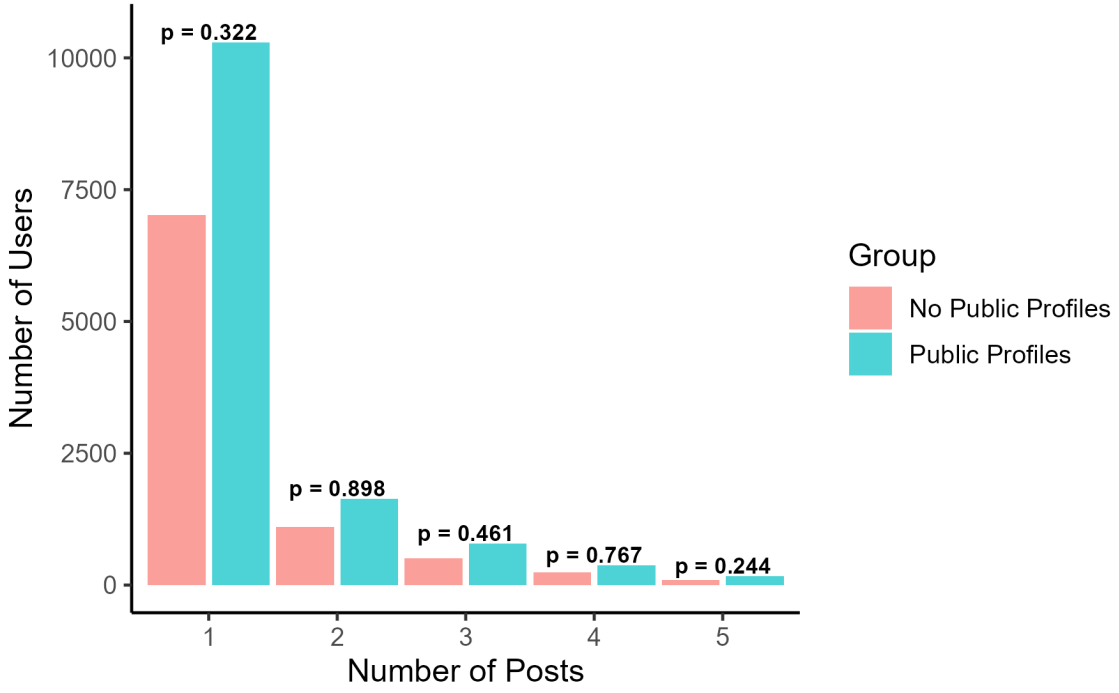
A Online Appendix: Additional Tables and Figures

Figure A.1: Distribution of Toxicity Scores



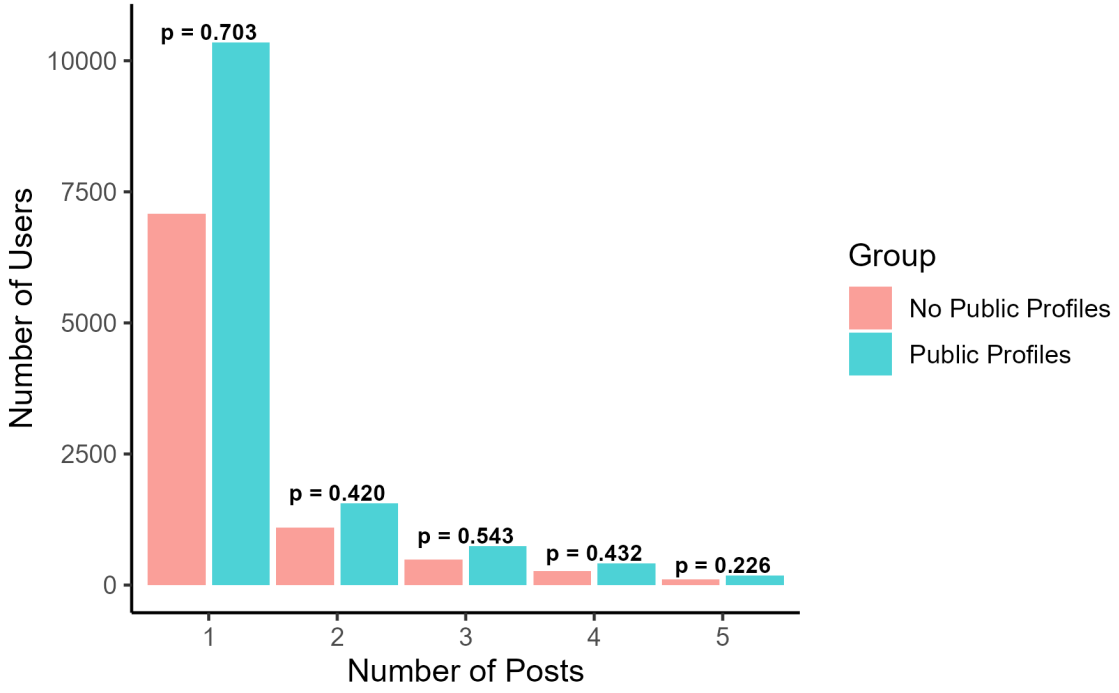
This figure illustrates the distribution of estimated toxicity scores for a collection of words, calculated via their cosine similarity to a reference "toxic" vector. The x-axis shows each word's toxicity score, while the y-axis (in log scale) represents the frequency of that word's usage. Color coding is applied based on thematic categories displayed in the figure: green for politics (e.g., 政治 [Politics], 総理 [Prime Minister]), blue for economy (e.g., 経済 [Economy], 財政 [Finance]), purple for repairs (e.g., 故障 [Breakdown], プリンター [Printer]), and red for hateful or offensive terms (e.g., バカ [Idiot], 死ね [Die], クズ [Trash]).

Figure A.2: Histogram of Posting Frequency (Public vs. No Public): The Introduction of the Like-button System



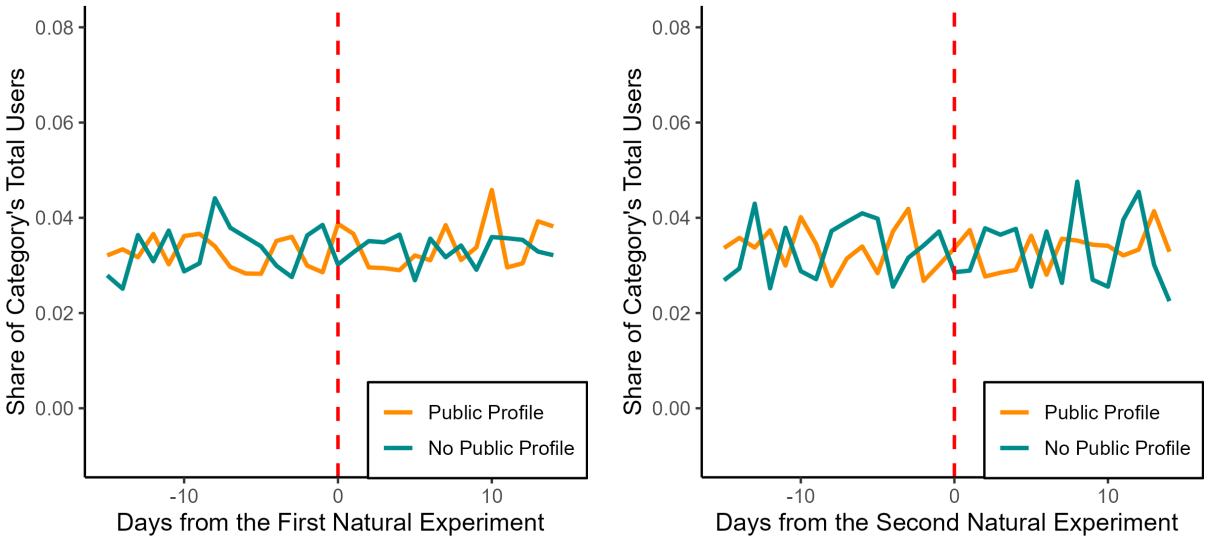
This figure illustrates how many users posted 1, 2, 3, 4, or 5 times during our target periods for the first natural experiment (i.e., the introduction of the like-button system). This is broken down into those with Public Profiles (in turquoise) and those without (in coral). Along the x-axis are the possible numbers of posts (1–5), while the y-axis shows the total number of users in each group. The p-values above each pair of bars come from two-proportion z-tests comparing the proportion of Public versus No Public users for that specific posting-frequency category.

Figure A.3: Histogram of Posting Frequency (Public vs. No Public): The Introduction of Rating-based Monetary Rewards



This figure illustrates how many users posted 1, 2, 3, 4, or 5 times during our target periods for the second natural experiment (i.e., the introduction of the rating-based monetary rewards). This is broken down into those with Public Profiles (in turquoise) and those without (in coral). Along the x-axis are the possible numbers of posts (1–5), while the y-axis shows the total number of users in each group. The p-values above each pair of bars come from two-proportion z-tests comparing the proportion of Public versus No Public users for that specific posting-frequency category.

Figure A.4: Number of Logged-in Users Before and After System Changes (Like-button Introduction and Rating-based Monetary Rewards)

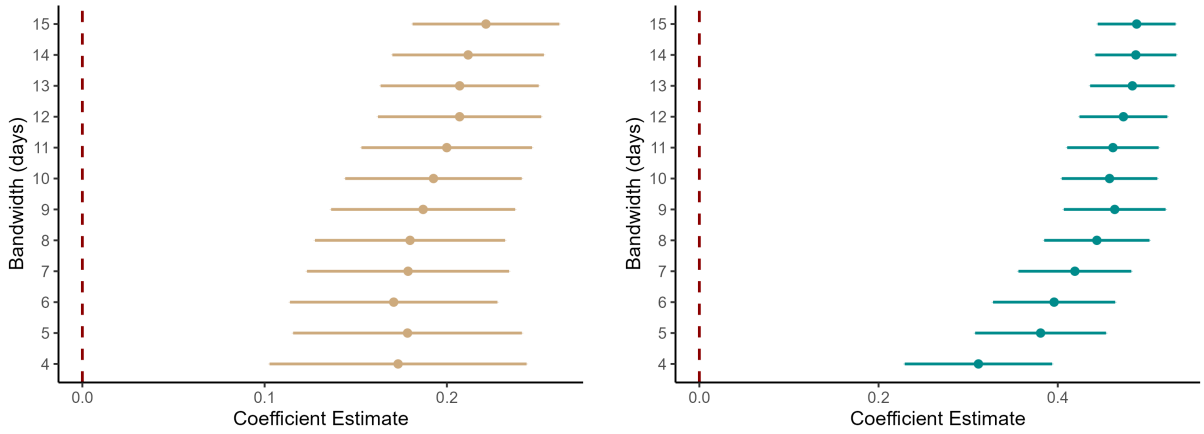


(a) The Introduction of the Like-button

(b) The Introduction of Monetary Rewards

Note: These figures illustrate the changes in the proportion of logged-in users by treatment status. We focus on the 15-day window before and after each system change. Each line shows the share of logged-in users relative to the total user base during the observed period. "Public Profile" indicates users who maintain a public SNS profile outside the platform, while "No Public Profile" represents those who do not. The x-axis denotes days relative to the treatment date.

Figure A.5: Robustness Checks: Different Bandwidths in the DiD Approach



(a) The Introduction of the Like-button

(b) The Introduction of Monetary Rewards

Note: This figure presents point estimates and 95% confidence intervals from DiD regressions designed to assess how the introduction of a platform policy affects toxicity scores. The regressions are based on the following specification: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\text{PublicAccount}_i + \delta\text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. $\text{ToxicScore}_{a,i,q,d}$ measures the similarity of an answer a, written by individual i in question thread q on day d, to content previously removed for toxicity. PublicAccount indicates whether the user's profile is public, and After_d is a post-policy indicator. The analysis is conducted using windows of 4 to 15 days around the policy introduction date. Orange markers represent estimates based on the introduction of the like-button while green markers represent estimates based on the introduction of monetary incentives.

Table A.1: A Comparison of Major Q&A Sites

Website Name	**# of Users**	**Year Est.**	**Host Country**	**Topic**	**Reputation Sharing**	**Monetary Incentives**	**Remarks**
Reddit	57 million daily active users as of December 2022	2005	USA	General	Yes	Yes	Known for its vast array of communities ("subreddits"). Popular for niche interests.
Quora	Over 300 million monthly visits as of 2023	2009	USA	General	Yes	No	Users can upvote the best answers; follow topics, questions, and people to build a personalized feed.
Stack Overflow	100 million monthly visits, 23 million registered users as of Nov 2022	2008	USA	Programming	Yes	No	A platform for programmers to ask and answer questions. Known for its extensive user base and community resources.
Yahoo! Answers	No longer exists, but 5.5 million monthly (US) visits as of Nov 2015	2005	USA	General	Yes	Yes	Yahoo! used comScore stats in Dec 2006 to proclaim Yahoo! Answers the leading Q&A site on the web.
Our Target Site	Over 70 million visits per year as of 2023	2000	Japan	General	No ↓ Yes (June 2018)	No ↓ Yes (Sep 2018)	One of the most popular Q&A sites in Japan, covering everything from politics to entertainment.

Note: Information is from Wikipedia. Reddit: https://en.wikipedia.org/wiki/Reddit; Quora: https://en.wikipedia.org/wiki/Quora; Stack Overflow: https://en.wikipedia.org/wiki/Stack_Overflow; Yahoo! Answers: https://en.wikipedia.org/wiki/Yahoo!_Answers. Note that Yahoo! Answers was discontinued by Yahoo on May 4, 2021, and is no longer available.

Table A.2: Summary Statistics of User Characteristics

	Panel A: Sample for the First Treatment					
	Overall		Pre-treatment		Post-treatment	
Variable	Mean	SD	Pre-Mean	Pre-SD	Post-Mean	Post-SD
User's Age	35.984	5.042	35.997	5.040	35.972	5.043
The Ratio of Male User	0.716	0.451	0.720	0.449	0.713	0.452
Months of User Experience	87.507	36.609	87.422	36.829	87.588	36.400

	Panel B: Sample for the Second Treatment					
	Overall		Pre-treatment		Post-treatment	
Variable	Mean	SD	Pre-Mean	Pre-SD	Post-Mean	Post-SD
User's Age	35.955	5.013	35.992	5.026	35.918	5.000
The Ratio of Male User	0.720	0.449	0.722	0.448	0.717	0.451
Months of User Experience	87.462	36.657	87.916	36.500	87.019	36.805

Note: This table presents summary statistics of user characteristics. "User's Age" indicates the average age of registered users. "The Ratio of Male User" indicates the ratio of male users. "Months of User Experience" indicates how many months users have been on the site since they first registered. Panel A focuses on the sample for the first treatment, while Panel B focuses on the sample for the second treatment. "Overall" contains a full sample; "Pre-treatment" contains a sample before the treatment; "Post-treatment" contains a sample after the treatment.

Table A.3: Alternative Outcomes: The Impact of Introducing the Like-button System and Rating-based Monetary Rewards

	Dep Var: Deleted Answer			
	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-like-button Dummy	0.0099**	0.0097**		
	(0.0042)	(0.0042)		
Having a Public Profile × Post-monetary-reward Dummy			0.0152*	0.0152*
			(0.0079)	(0.0079)
R^2	0.20990	0.21054	0.29342	0.29382
Observations	31,472	31,472	32,207	32,207
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects		✓		✓

Notes: This table presents regression results based on the following equation: $\text{Deleted}_{s,i,q,t,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is a binary variable indicating whether the answer was deleted by the platform. Columns (1) and (2) include all answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.4: Spillover Effects: Do Logged-in Treated Users and Their Toxic Content Affect Toxicity of Control Users?

	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	Share of Logged-in Public Account User	Toxic Content Share	Share of Logged-in Public Account User	Toxic Content Share
	(1)	(2)	(3)	(4)
The Share of Logged-in Public Account User in Threads (SHARE PUBLIC)	-0.0269		0.0164	
	(0.0224)		(0.0216)	
SHARE PUBLIC × Post Dummy	0.0094		-0.0209	
	(0.0291)		(0.0296)	
The Share of Toxic Content in Threads (SHARE TOXIC)		0.0109		-0.0066
		(0.0297)		(0.0301)
SHARE TOXIC × Post Dummy		-0.0484		0.0061
		(0.0402)		(0.0405)
R^2	0.36696	0.36694	0.34951	0.34947
Observations	12,468	12,468	12,531	12,531
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects	✓	✓	✓	✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta\,\text{ToxicExposure}_{a,i,q,d} + \delta\,\text{After}_d \times \text{ToxicExposure}_{a,i,q,d} + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score is estimated by a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1) and (2) contain all answers posted by control users (having no public social media accounts) within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted by control users (having no public social media accounts) within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.

Table A.5: Balance Tests for User Characteristics: The Like-button Introduction

	Dep Var: User Characteristics			
	Age	Gender	Tokyo Residence Dummy	Experience Years
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-like-button Dummy	-0.1161	0.0035	-0.0060	0.0091
	(0.1204)	(0.0112)	(0.0121)	(0.2963)
R^2	0.07499	0.07079	0.07650	0.07772
Observations	30,485	30,485	30,485	30,485
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓

Notes: This table presents regression results based on the following equation: $\text{Outcome}_{s,i,q,t,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is a user characteristic variable. The list of user characteristics is user age, gender, a Tokyo dummy for a residentail area, and experience years in a platform. Columns (1) to (4) contain all answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.6: Balance Tests for User Characteristics: The Monetary Reward Introduction

	Dep Var: User Characteristics			
	Age	Gender	Tokyo Residence Dummy	Experience Years
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-monetary-reward Dummy	-0.0325	-0.0047	-0.0138	0.3382
	(0.1338)	(0.0124)	(0.0134)	(0.3254)
R^2	0.07400	0.07774	0.07452	0.07519
Observations	30,640	30,640	30,640	30,640
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓

Notes: This table presents regression results based on the following equation: $\text{Outcome}_{s,i,q,t,d} = \alpha + \beta\text{PublicAccount}_i + \delta\text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is a user characteristic variable. The list of user characteristics is user age, gender, a Tokyo dummy for a residentail area, and experience years in a platform. Columns (1) to (4) contain all answers posted two weeks before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** p<0.01, ** p<0.05, * p<0.1.

Table A.7: Controlling for Order Effects: The Impact of Introducing the Like-button System and Rating-based Monetary Rewards

	Dep Var: Toxicity Score How Similar to Answers Removed by the Platform			
	Introduction of Like-button System		Introduction of Rating-based Monetary Reward	
	(1)	(2)	(3)	(4)
Having a Public Profile × Post-like-button Dummy	0.2215***	0.2179***		
	(0.0205)	(0.0207)		
Having a Public Profile × Post-monetary-reward Dummy			0.4881***	0.4838***
			(0.0221)	(0.0225)
R^2	0.28662	0.29601	0.32135	0.33504
Observations	30,485	30,485	30,640	30,640
Question fixed effects	✓	✓	✓	✓
Date fixed effects	✓	✓	✓	✓
Hour of the Day fixed effects		✓		✓
Order Effect fixed effects		✓		✓

Notes: This table presents regression results based on the following equation: $\text{ToxicScore}_{a,i,q,d} = \alpha + \beta \text{PublicAccount}_i + \delta \text{After}_d \times \text{PublicAccount}_i + \gamma' X_i + \theta_q + \zeta_d + \varepsilon_{a,i,q,d}$. The dependent variable is the toxicity score that is the similarity to answers removed by the platform due to their toxicity. The score is estimated by a word-embedding model. The score outcome is standardized to have a mean of zero and a standard deviation of one. Columns (1) and (2) include all answers posted within 15-day windows before and after the first natural experiment (i.e., the introduction of the like-button system). Columns (3) and (4) contain all answers posted within 15-day windows before and after the second natural experiment (i.e., the introduction of rating-based monetary incentives). In all specifications, the number of characters in a post is included as a control variable. Standard errors clustered at a question thread ID are reported in parentheses. *** $p<0.01$, ** $p<0.05$, * $p<0.1$.