Bao, Ying; Liu, Jessie

Working Paper

Spiral of Silence: Polarizing Content Creation through Moderating Toxicity

CESifo Working Paper, No. 12008

Provided in Cooperation with:
Ifo Institute – Leibniz Institute for Economic Research at the University of Munich

Suggested Citation: Bao, Ying; Liu, Jessie (2025) : Spiral of Silence: Polarizing Content Creation through Moderating Toxicity, CESifo Working Paper, No. 12008, CESifo GmbH, Munich

This Version is available at:
https://hdl.handle.net/10419/324999

CES ifo Working Papers

12008
2025

July 2025

Spiral of Silence: Polarizing Content Creation through Moderating Toxicity

Ying Bao, Jessie Liu

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

Spiral of Silence:
Polarizing Content Creation through Moderating Toxicity*

Ying Bao
University of Illinois Urbana-Champaign

Jessie Liu
Johns Hopkins University

July, 2025

Abstract

This paper investigates how content moderation affects content creation in an ideologically diverse online environments. We develop a model in which users act as both creators and consumers, differing in their ideological affiliation and propensity to produce toxic content. Affective polarization - users' aversion to ideologically opposed content - interacts with moderation in unintended ways. Even ideologically neutral moderation that targets only toxicity can suppress non-toxic content creation, particularly from ideological minorities. Our analysis reveals a content-level *externality*: when toxic content is removed, non-toxic posts gain exposure. While majority-group creators sometimes benefit from this exposure, they do not internalize the negative spillovers, i.e., increased out-group hostility toward minority creators. This discourages minority expression and polarizes the content supply, ultimately leaving minority users in a more ideologically imbalanced environment: a mechanism reminiscent of the "spiral of silence." Modeling creation as a strategic response to moderation, we underscore the importance of eliciting whether user engagement reflects toxicity or ideological disagreement in guiding platform governance.

Keywords: Content Moderation, Platform Design, Polarization, Toxicity

*The authors are listed alphabetically. We are grateful for the feedback received from participants at POMS, IIOC, the Marketing Science Conference, the CESifo Summer Institute on Digital Platforms, and seminars at the University of Washington and University of Illinois at Urbana-Champaign. We also thank Rafael Jiménez-Durán, Peter Landry, Alexei Makarin, Cameron Martel, Marcel Preuss, Mengze Shi, Shubhranshu Singh, Mateusz Stalinski, Michael Zhang, Yanhui Wu, and Pinar Yildirim for helpful conversations. All errors are our own.

1 Introduction

> "The opinion of only part of the population seemed to be the opinion of all and everybody, and exactly for this reason seemed irresistible to those who were responsible for this deceptive appearance."
>
> — Alexis de Tocqueville, *L'Ancien Régime et la Révolution* (1856, p. 259)

Toxicity is a pervasive challenge on social media platforms, affecting millions of users daily. According to the Pew Research Center, 4 in 10 Americans have faced online harassment, and 71% support tighter platform rules on toxic content.[1] Much of the existing literature on content moderation asks a demand-side question: How do users react when speech is suppressed? Prior work has shown that content removal can trigger backlash (Jhaver et al., 2019), reduce engagement (Jiménez Durán, 2021), or shift user composition (Liu et al., 2022). While these studies have advanced our understanding of moderation's impact on content consumption, they often overlook a supply-side question: How does the *anticipation* of moderation shape what speech exists in the first place?

Our study takes this supply-side perspective as its starting point. This focus is increasingly vital in today's social media landscape: moderation policies are no longer perceived as occasional or exceptional—they are institutionalized, expected, and often legally mandated (Andres and Slivko, 2021). As platforms implement preemptive tools such as automated flagging and removal, and as creators adapt to avoid penalties or backlash, anticipatory behavior becomes more relevant than reactionary responses alone. In this regime, it is crucial to understand not only which content gets removed, but also *which* content is deterred from being created at all.

A key motivation for our approach comes from a puzzling empirical pattern: opposite moderation policies can produce similar outcomes. When Reddit banned several toxic communities during "The Great Ban" in 2020, some users doubled their posting activity (Cima et al., 2024). Conversely, when Twitter relaxed its rules in 2022, hate speech also nearly doubled (Hickey et al., 2023). These two cases illustrate that moderation does more than remove harmful content: it changes the entire environment in which users create content. On social media, users are not merely passive consumers of content, they are also strategic creators who weigh the risks and rewards of content creation.

Despite growing attention to polarization, most moderation research conflate ideological division with harmful content, overlooking the distinct incentives and consequences that arise from ideology-

[1] https://www.pewresearch.org/internet/2021/01/13/the-state-of-online-harassment/

based versus toxicity-based preferences. This conflation obscures a key tension: content that is non-toxic in language may still provoke severe backlash when it violates ideological norms. For example, TikTok influencer Leo Skepi, with over 4 million followers, faced severe backlash after stating that brands should not be blamed for not carrying all sizes.[2] Disentangling these two types of preferences is therefore conceptually important, not only to understand how users engage with content, but also how they decide *what* to create. Our paper provides a formal framework to do so. Specifically, we model user preferences along two distinct dimensions: (i) a *vertical* preference capturing aversion to overtly toxic content, and (ii) a *horizontal* preference capturing aversion to ideologically misaligned content. This distinction allows us to characterize the conditions under which content moderation, by shifting visibility and engagement dynamics, can either suppress or stimulate content creation across ideological lines—even when the total volume of toxic content is reduced.

In this paper, we develop a model in which users act as both consumers and (potential) creators of content. They differ along two dimensions: their ideological position and their propensity to produce toxic content. Content creation is driven by a combination of intrinsic motivation and external engagement—whether positive or negative —both of which depend on the expected visibility of the content. Our model yields three key predictions. First, we show that it is not content removal per se, but the anticipation of moderation, that shapes creation incentives. When readers strongly prefer ideologically aligned content and creators are sensitive to negative feedback, moderation can undermine its own goals: toxic users may be motivated to create more content when lower exposure reduces the likelihood of backlash against their toxicity, while non-toxic users may create less in response to hostility from the ideological out-group following increased visibility. In such settings, moderation can ironically reward the group it aims to suppress while discouraging the group it seeks to protect.

Second, our model highlights the importance of distinguishing between two aspects of polarization. The first is *ideological imbalance*: an uneven distribution of users across ideology groups. The second is *affective polarization*: the degree to which users are averse to content that conflicts with their ideology, regardless of its tone or intent (Iyengar et al., 2019). This affective dimension plays a critical role in shaping how users engage with content: when it is strong, even civil posts may trigger backlash simply because they come from the "other side." Recent experimental studies also confirm this growing trend: users tend to evaluate otherwise identical content more

[2]https://time.com/6965324/leo-skepi-tiktok-clothing-size

negatively when it comes from ideological out-group members (Wuestenenk et al., 2025). These behavioral patterns align with our finding that, when one ideological group dominates the platform and affective polarization is strong, moderation can unintentionally amplify majority voices while suppressing minority expression—even when both groups host identical share of toxic users. This challenges the assumption that moderation is ideologically neutral, even if it targets only toxicity. In particular, while cross-cutting exposure may be effective in low (affective) polarization settings, it can be counterproductive when polarization is already high. Platforms should focus on measuring and reducing affective polarization itself, before attempting to redistribute attention across ideological lines. This could involve, for example, encouraging norm-based engagement or emphasizing shared values.

Third, we extend these insights to consumer welfare. Moderation redistributes welfare in asymmetric ways: while all consumers benefit from reduced exposure to toxic content (a universal gain), the composition of what remains skews toward content from the majority group (a polarizing effect). Hence majority readers are exposed to more ideologically aligned content whereas the minority users encounters the opposite. This asymmetry widens welfare inequality across ideological groups. These results have direct implications for policy frameworks such as the European Union's Digital Services Act (DSA), which emphasizes fairness and transparency in content moderation. Our findings suggest that such frameworks must move beyond static notions of fairness that focus only on what content is removed or demoted. Instead, they should account for the anticipatory feedback loops induced by moderation. Without accounting for the distinct roles of ideology and toxicity in user behavior, moderation policies may unintentionally reinforce polarization and marginalize under-represented groups.

Our results carry important implications for platform governance and the creator economy. Platforms such as Wattpad or YouTube, whose value depends heavily on sustained creator participation, face a fundamental trade-off: how to disentangle toxicity from ideological disagreement in order to mitigate the externalities of the former on creation shaped by the latter. This trade-off becomes especially acute when the user base is ideologically imbalanced and affective polarization is high. While platforms may not explicitly optimize for creator welfare, reader utility, or total engagement, our model maps the tensions they face under each objective. In particular, moderation policies that maximize positive engagement can unintentionally reinforce ideological imbalance if they fail to account for the externalities embedded in a polarized content supply. From a policy perspective, our findings echo the "spiral of silence" literature in social psychology (Noelle-Neumann,

1974), illustrating how visibility regimes, rather than speech restrictions alone, can marginalize minority voices. Our model also offers a theoretical foundation for recent efforts to design moderation mechanisms that distinguish toxicity from ideological disagreement (Twitter, 2021). However, we highlight a critical caveat: unless such mechanisms are incentive-compatible and elicit truthful reporting, flag-based moderation may institutionalize a new form of silence: not because content is genuinely harmful, but because it is unpopular with dominant groups and thus more likely to be mislabeled as "toxic."

Our paper contributes to three strands of literature. First, it advances research on the negative consequences of social media consumption. Prior work has documented that exposure to hate speech increases offline hate crimes (Andres and Slivko, 2021; Müller and Schwarz, 2021, 2023), while curation algorithms on platforms like Facebook and Twitter can amplify trolling, polarization, and echo chambers (Cinelli et al., 2021; Levy, 2021; Bondi et al., 2025; Pei and Mayzlin, 2024). Although previous studies have recognized the harms of either explicitly harmful content (vertical preference) or ideology-based backlashes (horizontal preference), these forces are often treated in isolation. In practice, they are two aspects manifested in the same content. By explicitly distinguishing between ideology-based and toxicity-based preferences, our paper provides a framework to understand how these two dimensions jointly shape online content creation and how moderation policies may unintentionally shift their incentives.

Second, we make a conceptual contribution to the literature on content moderation by emphasizing the dual role of users as both content consumers and strategic content creators. Existing work has shown that moderation policies can reduce audience engagement with hateful content (Thomas and Wahedi, 2023) and lead to lower hate-content production online as well as offline harm (Andres and Slivko, 2021; Jiménez Durán et al., 2024). Meanwhile, these interventions have been shown, in some cases, to reinforce echo chambers and reduce overall engagement (Huang et al., 2024). However, these average effects often obscure heterogeneity in user responses. Our study builds on this foundation by endogenizing content creation decisions across different user types. Rather than examining only how users respond to moderated content, we focus on how the *anticipation* of moderation, through its effects on visibility and expected engagement, reshapes the supply of content. This perspective helps reconcile some seemingly divergent empirical findings on how moderation affects content creation: what may appear as null or negative effects on average can, in fact, emerge from offsetting behavioral changes across users with different ideological identities and toxicity levels.

Third, we formally identify the role of affective polarization in a core marketing context: social media engagement. While affective polarization has been extensively studied in political science (Iyengar et al., 2019; Druckman and Levendusky, 2019), its implications for marketing remain underexplored (Godes et al., 2019). As partisan, racial, and religious identities increasingly converge, individuals are more likely to react emotionally to ideologically opposing content (Iyengar et al., 2019). The rise of partisan media further reinforces these group identities, making individuals more sensitive to perceived in-group and out-group cues, even when their core beliefs remain unchanged (Lelkes et al., 2017). This distinction is especially relevant in digital marketing, where identity signaling and perceived group affiliation are often inseparable from content consumption and creation. For example, affective polarization can intensify backlash to brand activism (Homroy and Gangopadhyay, 2023), affect trust in corporate messaging across ideological lines (Liaukonytė et al., 2023), or complicate influencer partnerships when perceived affiliations diverge from audience values (Schad, 2023). Our framework offers a tractable approach for marketing scholars to model these dynamics and examine how polarization interacts with platform design, ultimately affecting consumer engagement and brand outcomes.

The rest of the paper is organized as follows. Section 2 introduces the model and equilibrium concept. In Section 3, we assess the impact of moderation. In Section 4, we extend the model by exploring different incentives for content creation. Section 5 discusses several model variations, and Section 6 concludes with policy and managerial implications.

2 Model

A social media platform hosts a population of users with a total mass of 1. Each user plays a *potential* dual role as both a content creator and a content consumer—producing content for others to engage with, while also reading and interacting with content created by others. Users differ in two dimensions: their ideology type $i \in \{A, B\}$ and toxicity type $t \in \{T, NT\}$. The ideology type i reflects the view points they lean toward, such as Democrat vs. Republican, or pro-vaccine vs. vaccine-hesitant. The toxicity type reflects their propensity to post toxic ($t = T$) or non-toxic ($t = NT$) content. We assume that each user can create up to one piece of content that matches with their type.

We introduce a parameter $\delta \in [0, \frac{1}{2}]$ to capture the degree of *ideological imbalance* in the population of platform users. It measures the degree of asymmetry in ideological group size.

Specifically, the total shares of users with ideology A and B are given respectively by $(\frac{1}{2}+\delta)$ and $(\frac{1}{2}-\delta)$. In other words, ideology A group is assumed to represent the majority group. This assumption is innocuous, as we remain agnostic about which group is more prone to toxicity—both are treated symmetrically by construction. Let $x \in (0,1)$ denote the overall mass of toxic users in the population. Let $\tau_A \in (0,1)$ and $\tau_B = 1-\tau_A$ denote the share of toxic users who belong to ideological groups A and B, respectively. For example, when $\tau_A = 1$ and $\tau_B = 0$, all toxic users are from group A; when $\tau_A = 0$, all toxic users are from group B. Table 1 below summarizes the resulting population mass of each user type:

	Toxic (T)	Non-toxic (NT)	**Total**
Ideology A	$x \cdot \tau_A$	$(\frac{1}{2}+\delta) - x \cdot \tau_A$	$\frac{1}{2}+\delta$
Ideology B	$x \cdot \tau_B$	$(\frac{1}{2}-\delta) - x \cdot \tau_B$	$\frac{1}{2}-\delta$
Total	x	$1-x$	1

Table 1: Mass of User Type by Ideology and Toxicity

In the subsequent sections, we use the 2-tuple $(i,t) \in \Theta \equiv \{A,B\} \times \{T, NT\}$ to denote a user's type. We use $\lambda(i,t)$ to denote the population share of a type (i,t) user.

2.1 Content Consumption

A reader of type (i,t) on the platform is exposed to content generated by other users. Their utility from consuming a particular piece of content depends on two key factors: whether the content aligns with the reader's ideological orientation and whether it contains toxic elements. We represent the utility of a reader r of type (i,t) from consuming content created by a user of type (i',t') as:

$$U^r(it, i't') = \underbrace{\alpha \cdot H(i,i')}_{\text{horizontal (ideology-based)}} + \underbrace{(1-\alpha) \cdot V(t')}_{\text{vertical (toxicity-based)}} + \varepsilon_r, \tag{1}$$

where

$$H(t,t') = \begin{cases} 0 & \text{if } i = i' \\ -1 & \text{if } i \neq i' \end{cases}, \qquad V(t') = \begin{cases} 0 & \text{if } t' = NT \\ -1 & \text{if } t' = T \end{cases}.$$

This utility specification[3] captures two dimensions of content evaluation commonly observed on social media platforms. The first, referred to as the *horizontal* dimension ($H(\cdot)$), reflects the tendency for users to disfavor ideologically misaligned content, which is consistent with evidence that users are more receptive to viewpoints that match their own (Kozyreva et al., 2023). The second, the *vertical* dimension, reflects a general aversion to harmful or toxic content, independent of ideological alignment. Note that our utility specification suggests that a user's own toxicity type $t \in \{NT, T\}$ does not affect their utility from consuming content, i.e., $U^r(it, i't') = U^r(i, i't')$. This is because users judge others' content based on its ideology and tone, regardless of their own posting behavior. In other words, even users who themselves create toxic content may still exhibit aversion to others' toxicity, particularly when the content is ideologically misaligned.

The parameter $\alpha \in [0, 1]$ calibrates the relative weight users place on ideological alignment versus content toxicity. One can also interpret α as the degree of aversion to the opposing ideology. When α is close to 1, users prioritize ideological alignment over toxicity concerns. This formulation resonates with the political science literature on "affective polarization": the increasing animosity between the parties, even in the absence of substantive ideological divergence. As discussed by Iyengar et al. (2019), affective polarization stems from partisanship functioning as a social identity, where individuals categorize others into a favored in-group and a disfavored out-group, independent of policy-specific disagreement (p. 130). Finally, $\varepsilon_r \sim U[-1, 1]$ denotes users' idiosyncratic utility shock from consuming content.

Upon consuming content created by others, a reader can decide whether to engage with the content by liking or disliking the content. The like and dislike decisions do not necessarily mean the like button or dislike button. Instead, we use them to represent all positive and negative engagement with the posts, including reactive emojis and comments directed towards the posts. Research suggests that social media users are generally more inclined to express positive feedback, such as "likes," rather than negative feedback, such as "dislikes." This tendency is influenced by platform design, such as the prominent display of like buttons, and psychological factors, including the drive for social validation (Stsiampkouskaya et al., 2023). Thus, we assume that a reader will like a post if the utility from consuming it is greater than 0, i.e. $U^r \geq 0$, whereas they will dislike a post only if the utility is below a threshold, i.e., $U^r \leq -\gamma$. Here, $\gamma \in (0, 1)$ can be considered as the relative cost of negative engagement: the higher γ is, the less likely a reader will dislike the

[3]The additive structure is a simplifying assumption for analytical tractability and conceptual clarity. It provides a micro-foundation for the probabilistic model of user engagement later described in Table 2.

content. Figure 1 below illustrates readers' engagement pattern induced by their utility $U^r(\cdot)$.

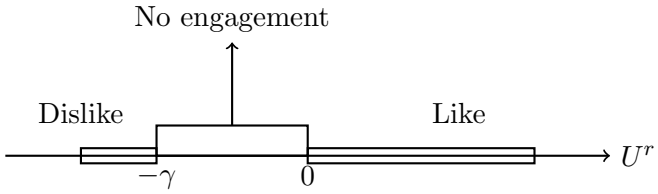

Figure 1: User Engagement Based on $U^r(\cdot)$

For simplicity, we normalize γ to $\frac{1}{2}$ in the main model.[4] As discussed above, the readers' toxicity type does not affect their consumption utility, i.e., $U^r(it, i't') = U^r(i, i't'), \forall t \in \{NT, T\}$. Thus, we use $P_l(r = i, c = i't')$ and $P_{dl}(r = i, c = i't')$ to denote respectively the probability of positive and negative engagement of a reader r consuming a post created by a creator c. Given the utility specification, Table 2 summarizes these probabilities in different cases.

$P_l(r,c)$		**Creator**			
		A, NT	A, T	B, NT	B, T
Reader	A	$\frac{1}{2}$	$\frac{\alpha}{2}$	$\frac{1-\alpha}{2}$	0
	B	$\frac{1-\alpha}{2}$	0	$\frac{1}{2}$	$\frac{\alpha}{2}$

(a) Positive Engagement

$P_{dl}(r,c)$		**Creator**			
		A, NT	A, T	B, NT	B, T
Reader	A	$\frac{1}{4}$	$\frac{3-2\alpha}{4}$	$\frac{1+2\alpha}{4}$	$\frac{3}{4}$
	B	$\frac{1+2\alpha}{4}$	$\frac{3}{4}$	$\frac{1}{4}$	$\frac{3-2\alpha}{4}$

(b) Negative Engagement

Table 2: Positive (a) and Negative (b) Engagement Probabilities by Reader and Creator Type

Taken together, these probabilities suggest that users are more likely to positively engage with (like) non-toxic content created by users with the same ideology. Meanwhile, users are more likely to negatively engage with (dislike) toxic content from creators with the opposite ideology. These behavioral patterns also align with the empirical findings documented in Kozyreva et al. (2023).

2.2 Content Creation and Moderation

In addition to reading others' posts, users can also create content. Prior work suggests that content creation is motivated by both intrinsic utility and the desire for external recognition or engagement

[4]Note that this normalization does not qualitatively change our results as long as it is exogenously given between 0 and 1.

(Toubia and Stephen, 2013). We formalize the utility of content creation for a user (creator) of type (i,t) as follows:

$$U^c(i,t) = \underbrace{S(t)}_{\text{survival rate}} \cdot \underbrace{P_s}_{\text{visibility}} \cdot \underbrace{(L(i,t) + \omega DL(i,t))}_{\text{external engagement}} \underbrace{+\varepsilon_c}_{\text{intrinsic utility}}, \tag{2}$$

where $\varepsilon_c \sim U[-1,1]$ captures idiosyncratic or intrinsic motivations unrelated to external reactions. The first three terms jointly capture the expected utility from external engagement, shaped by the following three factors:

1. Content Survival ($S(t)$): Content must survive moderation to be visible. Let $\beta \in [0,1]$ denote the intensity of the platform's moderation policy, representing the share of toxic posts removed before exposure. We define survival probability as:

$$S(t) = 1 - \beta \cdot \mathbf{1}[t = T],$$

i.e., only toxic posts are subject to removal. This formulation allows for partial enforcement, reflecting regulatory limits driven by free speech concerns or technical feasibility (Dave, 2020; Carlson and Rousselle, 2020). A regime with full enforcement ($\beta = 1$) removes all toxic posts, whereas $\beta = 0$ implies no moderation. A post fails to go live with probability $1 - S(t)$, in which case the creator receives only the intrinsic utility.

2. Content Visibility (P_s): Once live, content competes for user attention. We assume the platform recommends one post per reader, chosen uniformly at random from all surviving content. The probability that a given post gets seen decreases with the total volume of surviving content. For analytical tractability, we model this *crowding* effect via a simple linear approximation[5]:

$$P_s = 1 - \sum_{i\in\{A,B\}, t\in\{T,NT\}} \lambda(i,t) \cdot S(t) \cdot P_c(i,t), \tag{3}$$

where $\lambda(i,t)$ denotes the population share of type (i,t) users (as shown in Table 1), and $P_c(i,t)$ is the expected share of content created by such users. This assumption reflects a well-documented empirical regularity: from the creator's perspective, an increase in others' content reduces the likelihood that their own post is seen. For instance, Hodas and Lerman

[5]This linear approximation not only ensures analytical tractability but also preserves the key feature of its probabilistic counterpart, $\frac{1}{\sum_{i\in A,B,t\in T,NT} \lambda(i,t)\cdot S(t)\cdot P_c(i,t)}$: it monotonically decreases with the total volume of content available on the platform.

(2012) find that the probability of retransmitting a tweet declines roughly in proportion to the number of incoming messages. Similarly, Facebook's ranking pipeline trims "thousands of candidate posts" to just a few hundred, necessarily reducing exposure as content supply grows (Lada et al., 2021). While stylized, this baseline assumption provides a broadly applicable starting point.[6] As long as $P_c(i,t) \in [0,1]$, we have $P_s \in [0,1]$.

3. User Engagement ($L(i,t) + \omega DL(i,t)$): If a post is seen, it may generate both positive (likes) and negative (dislikes) engagement from readers. These are calculated as expected responses from all reader types:

$$L(i,t) = \sum_{(i',t') \in \Theta} \lambda\left(i',t'\right) \cdot P_l\left(i',it\right) \tag{4}$$

$$DL(i,t) = \sum_{(i',t') \in \Theta} \lambda\left(i',t'\right) \cdot P_{dl}\left(i',it\right). \tag{5}$$

where P_l and P_{dl} denote, from the perspective of a type (i,t) creator, the probabilities that their content receives a like or dislike, respectively, from readers of type (i',t'). The parameter $\omega \in [-1,0)$ captures the discouraging effect of negative engagement on content creation. Equivalently, $-\omega$ represents the creator's aversion to negative engagement. This is consistent with empirical findings that negative feedback tends to suppress creator activity (Berger and Milkman, 2012; Our Mental Health, 2025). We explore the possibility that some users may derive positive utility from negative engagement (i.e., $\omega > 0$) in Section 4.

A user of type (i,t) will choose to create content if and only if: $U^c(i,t) > 0$.

2.3 Equilibrium Concept

We adopt the concept of rational expectation equilibrium (Grossman and Stiglitz, 1980; Moorthy, 1985), in which the expected share of users of each type who create content matches the actual probability of content creation in equilibrium. Formally, the following condition must hold for all user types:

$$P_c(i,t) = \Pr[U^c(i,t) > 0], \quad \forall i \in \{A,B\}, t \in \{T,NT\}. \tag{6}$$

In equilibrium, all creators share a common belief about $P_c(i,t)$, which enters the visibility term P_s and must be internally consistent with the realized outcome.

[6] Our framework can be extended to incorporate more complex curation rules, but we begin with this natural benchmark to highlight the key strategic trade-offs.

In the next section, we discuss how moderation shapes content creation of each types and their respective welfare. We provide the detailed equilibrium characterization in Appendix A.1.

3 Analysis

We begin our analysis with two benchmark cases: (i) no affective polarization ($\alpha = 0$) and (ii) affective polarization present ($\alpha > 0$) with a balanced ideological composition ($\delta = 0$). We then relax these constraints to explore the full parameter space, considering environments characterized by both affective polarization among users ($\alpha > 0$) and ideological imbalance between groups ($\delta > 0$). By comparing equilibrium outcomes in case (ii) to case (i), we isolate the effect of vertical, toxicity-based preferences from that of horizontal, ideology-based preferences. This comparison demonstrates the role of affective polarization in shaping creators' incentives under content moderation. Comparing the full model to case (ii) allows us to examine how the effects of both horizontal and vertical preferences are further amplified or mitigated when one ideological group dominates the platform. Throughout our analysis, we maintain the assumption that the share of toxic users is not too large ($x < \frac{3}{4}$). This restriction simplifies our analysis and helps us focus on equilibrium outcomes that are most relevant in practice, where moderation unambiguously reduces the equilibrium exposure of toxic content. In Appendix A.1, we show that there exists a unique equilibrium and provide the detailed equilibrium characterization in each cases. Proofs for all subsequent result and propositions are also included in the Appendix.

Result 1. ***(No Affective Polarization)*** *When $\alpha = 0$, content moderation unambiguously increases the equilibrium level of content creation from both ideological groups. Specifically, $\frac{\partial P_c(i,NT)}{\partial \beta} > 0$ and $\frac{\partial P_c(i,T)}{\partial \beta} > 0$, but we have $\frac{\partial [S(T) \cdot P_c(i,T)]}{\partial \beta} < 0$, $\forall i \in \{A, B\}$. Furthermore, both total positive engagement and reader welfare monotonically increases with moderation intensity (β), i.e., $\frac{\partial TL}{\partial \beta} > 0$ and $\frac{\partial \mathbf{E}U^r(i)}{\partial \beta} > 0$, $\forall i \in \{A, B\}$.*

We provide the formal expressions of total likes (TL) and reader utility ($EU^r(i)$) in the Appendix. In this benchmark case where readers do not exhibit affective polarization ($\alpha = 0$), their engagement with the content depends solely on content toxicity. Consequently, from the creator's perspective, expected engagement is determined entirely by the toxicity level of their content, regardless of their ideological group. As a result, the effect of moderation on content creation is symmetric across ideological groups. Removing toxic content unambiguously increases the creation of non-toxic content by increasing their visibility. This increased visibility leads to more positive

engagement as users prefer non-toxic content, incentivizing non-toxic creators from both ideological groups to produce more. Interestingly, we also find that toxic users may increase their content creation. This occurs because they anticipate reduced backlash due to lower expected exposure under moderation, making their behavior increasingly driven by intrinsic utility rather than strategic considerations. Despite this, the overall volume of toxic content that *survives* moderation declines monotonically.

As a result, the positive effects of moderation extend beyond content creation to platform-wide outcomes. Total positive engagement increases with moderation intensity, as the higher visibility and growing volume of non-toxic content create a virtuous cycle that encourages further positive engagement. Moreover, reader welfare improves as moderation intensifies, since users benefit from both greater access to high-quality (non-toxic) content and reduced exposure to toxic content.

Overall, Result 1 suggests that, when readers exhibit only toxicity-based (vertical) preference rather than ideological-based (horizontal) preference, content moderation policy can generate a *win-win-win* outcome for creators, readers, and the platform as a whole. This seemingly intuitive finding carries significant implications for the design of content moderation policies. Wikipedia, for example, emphasizes factual verification over ideological alignment, demonstrating how the consistent removal of low-quality or toxic content can enhance user engagement and overall platform vitality.[7] More broadly, on platforms where users prioritize factual accuracy and content quality above ideological considerations (i.e., platforms characterized by low affective polarization, $\alpha \approx 0$), systematic moderation of harmful content or misinformation can yield substantial shared benefits.

Next we examine how content moderation affects equilibrium outcomes when readers exhibit some degree of affective polarization ($\alpha > 0$). In this case, readers demonstrate *in-group* validation and *out-group* hostility: they are more tolerant of toxic content from creators who share their ideological orientation while less forgiving of toxicity from opposing groups (as shown in Table 2). Because of this asymmetric preference, the impact of moderation on content creation is contingent on both the degree of affective polarization and the relative sizes of ideological groups. To isolate the effects of affective polarization, we first analyze a symmetric setting with equal-sized ideological groups ($\delta = 0$), where content creation incentives are identical across ideological groups, regardless of their toxicity composition, i.e., $P_c^*(A,t) = P_c^*(B,t)$ for all $t \in \{T, NT\}$. In this setting, any impact of moderation arises solely from affective polarization, as summarized in the following proposition.

[7] https://wikimediafoundation.org/news/2024/11/25/love-wikipedia-get-to-know-the-nonprofit-behind-it/

Proposition 1. ***(Impact of Affective Polarization)*** *When the two ideological groups are balanced ($\delta = 0$), but affective polarization ($\alpha > 0$) is strong and creators are sufficiently averse to negative engagement ($\omega \leq -\frac{1}{2}$), moderating toxic content can paradoxically reduce content creation by non-toxic users while increasing it among toxic users, across both ideological groups. Formally, $\forall i \in \{A, B\}$, we have*

$$\frac{\partial P_c(i, NT)}{\partial \beta} = \begin{cases} < 0 & \text{if } \alpha \in (\alpha_1(\omega), 1] \\ \geq 0 & \text{if } o/w \end{cases},$$

$$\frac{\partial P_c(i, T)}{\partial \beta} = \begin{cases} \geq 0 & \text{if } \alpha \in (0, \alpha_2(\omega)] \\ < 0 & \text{if } o/w \end{cases},$$

where $\alpha_1(\omega) \equiv \frac{2+\omega}{1-\omega}$ and $\alpha_2(\omega) \equiv \frac{-3\omega}{1-\omega}$. We have $\alpha_1(\omega) \leq 1 \leq \alpha_2(\omega)$ if $\omega \in [-1, -\frac{1}{2}]$ and $\alpha_2(\omega) < 1 < \alpha_1(\omega)$ if $\omega \in (-\frac{1}{2}, 0)$.

Proposition 1 shows that when affective polarization influences reader utility, moderating toxic content no longer necessarily increases creation by non-toxic users, as it does in Result 1. In fact, when affective polarization is sufficiently strong, non-toxic users reduce their content creation in equilibrium. This occurs when out-group hostility outweighs in-group validation—a dynamic amplified by increased visibility due to moderation. As moderation removes more toxic posts, non-toxic content is more likely to reach readers from both ideological groups. Conditional on exposure, users' aversion to negative engagement ($-\omega$) amplifies the harm of hostility relative to the benefit of validation, whether from in-group or out-group members; As α increases, out-group validation weakens while their hostility intensifies; once α exceeds a critical threshold $\alpha_1(\omega)$, the disutility from such negative engagement outweighs the benefits of positive engagement, i.e., $L(i, NT) + \omega DL(i, NT) < 0$. These factors initiate a negative *feedback loop* where increased visibility of non-toxic content, rather than facilitate a positive discourse, exposes creators to more negative engagements, which ultimately discourages their content creation.

The impact on toxic content creation presents an equally counterintuitive pattern that warrants separate examination. We identify two scenarios where toxic users paradoxically increase their content creation despite moderation risks: one driven by reduced backlash against content toxicity itself, and the other driven by reduced backlash stemming from ideological out-group hostility. First, when readers exhibit a low degree of affective polarization, that is, they care more about toxicity than ideological alignment: toxic users respond to stricter moderation by creating

more. The logic mirrors the benchmark case of no polarization ($\alpha = 0$): stricter moderation reduces expected exposure, thereby mitigating the anticipated backlash against toxicity itself. This leads toxic users to create content based more on intrinsic utility than on strategic considerations. Second, when creators have a strong aversion to negative engagement ($\omega < -\frac{1}{2}$), toxic users once again increase their content creation under stronger moderation, regardless of the level of affective polarization. This occurs because, even when readers prioritize ideological alignment over toxicity, reduced expected exposure softens backlash from the opposing group, thereby encouraging toxic users to create more. It is important to note that even though toxic users create more in these cases, the overall share of toxic content that *survives* moderation—$S(T) \cdot P_c(i, T)$—still decreases monotonically with the intensity of moderation (β). This implies that the direct effect of removal dominates the indirect effect from increased toxic creation.

Based on the discussion above, we have the following remark.

Remark 1. *When both affective polarization (α) and aversion to negative engagement ($-\omega$) are strong, moderation can inadvertently reward the group it seeks to suppress while discouraging the group it intends to protect.*

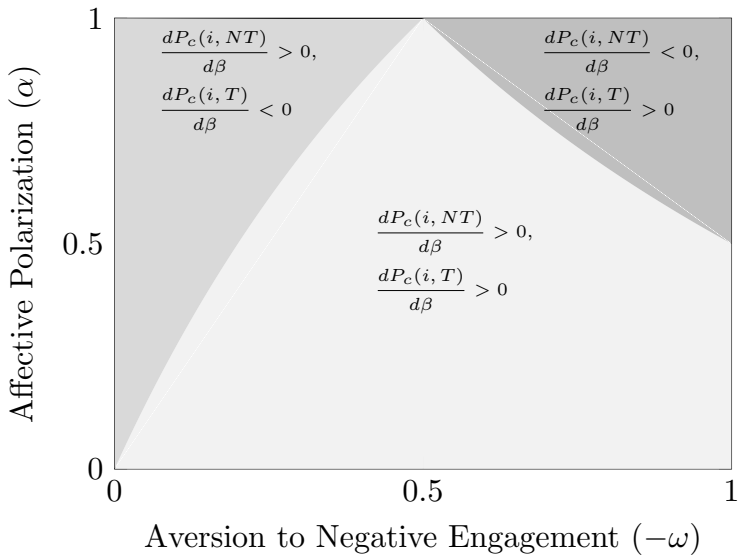

Figure 2: Effect of Moderation on Content Creation ($\delta = 0$)

Figure 2 illustrates how moderation influences content creation across different parameter regimes. When affective polarization (α) is low, moderation universally increases content creation by non-toxic users. At high α levels, if creators are relatively insensitive to negative feedback

(i.e., low $-\omega$), moderation can still achieve a desirable outcome: it suppresses toxic users while encouraging non-toxic ones. However, when both affective polarization and aversion to negative engagement are strong, our model reveals a perverse outcome: moderation incentivizes toxic users to create more, while discourages non-toxic users. Remarkably, these dynamics can arise solely from affective polarization ($\alpha > 0$), even in the absence of any ideological imbalance between groups ($\delta = 0$).

The policy implications are significant: effective moderation is not merely about removing toxic content, but about designing the right incentives. Without complementary measures, well-intentioned interventions risk reinforcing the very ideological divisions they aim to heal. In particular, our findings caution against blanket strategies such as promoting cross-cutting exposure in highly polarized environments; while promising in low-polarization settings, such exposure may be counterproductive and further intensify ideological backlashes when α is high. Instead, platform design should aim to reduce affective polarization itself, for instance, by encouraging norm-based engagement or highlighting common ground, rather than simply redistributing attention.

Next we turn to the case where both ideological imbalance ($\delta > 0$) and affective polarization ($\alpha > 0$) are present. It allows us to identify how unequal group sizes and preference for ideological alignment jointly shape the effects of content moderation. In the following discussion, we focus on how incentives for non-toxic creators shift relative to the two benchmark cases, given their central role in sustaining long-term content quality and user retention on the platform.

Proposition 2 (**Ideological Imbalance and Affective Polarization**). *For $\delta > 0$ and $\alpha > 0$, moderating toxic content affects the equilibrium content creation among non-toxic creators from ideological groups A and B differently. Specifically, three distinct regions emerge:*

- ***Universal Suppression**: If $\alpha > \bar{\alpha} = \frac{\omega+2}{(1-2\delta)(1-\omega)}$, moderation reduces non-toxic content creation from both groups, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} < 0, \frac{\partial P_c^*(B,NT)}{\partial \beta} < 0$.*
- ***Universal Empowerment**: If $\alpha < \underline{\alpha} = \frac{\omega+2}{(1+2\delta)(1-\omega)}$, moderation increases non-toxic content creation from both groups, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} > 0, \quad \frac{\partial P_c^*(B,NT)}{\partial \beta} > 0$.*
- ***Polarized Expression**: If $\underline{\alpha} \leq \alpha \leq \bar{\alpha}$, moderation amplifies polarization: majority group creates more whereas minority group creates less, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} > 0, \frac{\partial P_c^*(B,NT)}{\partial \beta} < 0$.*

Proposition 2 shows that when one ideological group dominates the population, moderation no longer affects creators uniformly across ideological lines. Instead, its impact depends critically on

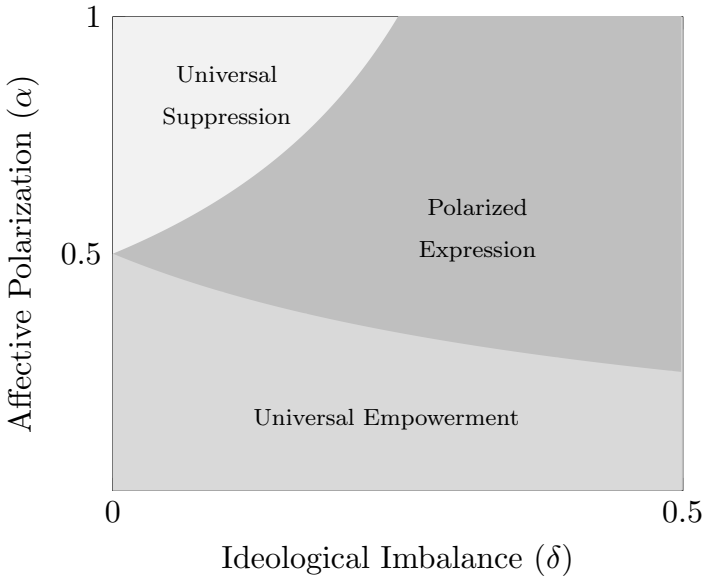

Figure 3: Effect of Moderation on Non-toxic Content Creation ($\omega = -1$)

the degree of affective polarization (α) and the extent of ideological imbalance (δ). As illustrated in Figure 3, three distinct equilibrium regions emerge. When affective polarization is weak ($\alpha < \underline{\alpha}$), we are in the "universal empowerment" region: moderation enhances the visibility of non-toxic content, and the benefits outweigh the costs for non-toxic creators across both groups. This logic echoes that of Result 1. At the other extreme, when affective polarization is strong ($\alpha > \bar{\alpha}$), we enter the "universal suppression" region: moderation triggers a negative feedback loop driven by intensified out-group hostility, ultimately reducing content creation from both sides. This rationale aligns with Proposition 1. Between these two extremes lies the "polarized expression" region, which arises under moderate affective polarization ($\underline{\alpha} \leq \alpha \leq \overline{\alpha}$). Here, moderation amplifies existing ideological imbalances: creators from the majority group are encouraged to produce more, while those from the minority group withdraw. The intuition is as follows: moderation removes toxic content, increasing the visibility of non-toxic posts from both groups. However, the resulting engagement dynamics differ sharply: majority creators anticipate net positive engagement, while minority creators expect intensified backlash from the larger opposing group, which can outweigh support from their own smaller base.

This final result from Proposition 2 echoes the "spiral of silence" effect, a well-documented phenomenon in the social psychology literature on public opinion formation (Noelle-Neumann, 1974). The term refers to the tendency of majority voices to dominate public discourse while

minority voices become silent, initiating "a spiraling process which increasingly establishes the [majority] opinion as the prevailing one" (Noelle-Neumann, 1974, p.44). This dynamic has far-reaching implications for the forecasting of political opinions, fashion trends, and the evolution of social norms. Traditionally, it is viewed as a psychological response to social pressure. However, our model shows that even well-intended content moderation can unintentionally accelerate this process. Specifically, moderation, rather than directly silencing the voices from minority groups, reshapes content visibility and creator incentives in a way that disproportionately discourages their creation. This mechanism also helps explain recent empirical findings that, while effective at reducing harmful material, moderation can disproportionately suppress or distort legitimate input from ideological or demographic minorities (Broniatowski et al., 2023; Haimson et al., 2021).

Taken together, these findings shed new light on a common debate: some conservative commentators attribute higher takedown rates to ideological "censorship," while some empirical studies suggest that right-leaning users may host more toxic users (Haimson et al., 2021). Our results offer an alternative perspective. Even when the share of toxic users that belong to each group is held constant, stricter moderation can *differentially* suppress or amplify non-toxic content across groups, driven entirely by supply-side responses to expected engagement. Thus, the empirical pattern in which one side "loses" more content need not reflect inherent toxicity; it may instead reflect how moderation policies interact with ideological imbalances and the incentives faced by non-toxic creators.

Building on Proposition 2, which shows how moderation differentially affects content creation across ideological groups, we now turn to its welfare implications. Rather than analyzing each group in isolation, we focus on the welfare inequality between two groups, i.e., $\mathbb{E}[U^r(A) - U^r(B)]$. This metric clarifies how moderation policies shape relative outcomes between ideological groups and influence the broader dynamics of polarization and perceived fairness.

For clarity and simplicity, in the following proposition, we assume an equal mass of toxic users across ideological groups, i.e., $\tau_A = \tau_B = \frac{1}{2}$. This is because when either τ_A or τ_B dominates, the result on welfare inequality is intuitive as the dominating group will bear most of the impact of content moderation. Moreover, this constraint helps disentangle the effects of toxicity and ideology by removing any built-in correlation between the two in user composition. In doing so, it sheds new light on the above common debate of whether there is ideological differences in toxicity. Our setup allows us to examine how ideological imbalance (δ) and affective polarization (α) shape moderation

outcomes, independent of any ideological differences in toxicity.[8]

Proposition 3. *(**Welfare Redistribution**) For $\alpha > 0$ and $\delta > 0$, stricter moderation (larger β) simultaneously (i) increases the platform's total positive engagement, i.e., $\frac{\partial TL}{\partial \beta} > 0, \forall \beta \in [0,1]$; and (ii) enlarges readers' welfare gap across two ideology groups. That is,*

$$\frac{\partial \mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta} > 0, \forall \beta \in [0,1].$$

Proposition 3 shows that while moderating toxic content increases overall positive engagement, it also redistributes reader welfare across ideological groups. The intuition naturally follows Proposition 2: although both groups benefit from reduced exposure to toxic content (a universal gain), the composition of what remains tilts toward content from the majority group (a polarizing effect). As a result, majority readers are exposed to more ideologically aligned content whereas the minority users face the opposite. This asymmetry drives an increasing slope in welfare inequality.

Once again, our findings from Propositions 2 and 3 underscore a fundamental challenge for platform governance: moderating offensive content without undermining ideological diversity. Viewed through a Coasean lens (Coase, 1960), moderating toxicity generates externalities that disproportionately burden non-toxic ideological minorities. When toxic content is removed, non-toxic posts gain greater exposure. Majority-group creators benefit from this shift but do not internalize the negative spillovers, namely, increased out-group hostility directed at minority creators. This reduces minority creation and polarizes content supply, leaving minority readers in a more ideologically imbalanced environment. To internalize this externality, platforms may benefit from designing mechanisms, such as flagging systems that distinguish between offensive content and ideological disagreement, that help disentangle toxicity aversion from affective polarization. In fact, some platforms have begun allowing users to categorize content as "toxic language," "misinformation," or "offensive ideology." For example, Twitter's Birdwatch (now Community Notes) program enables users to label and contextualize misleading or offensive tweets (Twitter, 2021).

However, such systems remain vulnerable to strategic misreporting. To address this distortion, platforms must implement incentive-compatible mechanisms that promote truthful reporting. Without such design, flag-based moderation risks institutionalizing a spiral of silence—not because content is *actually* harmful, but because it is unpopular with dominant groups. In such cases, the negative externalities of toxicity become increasingly difficult to disentangle from ideological

[8] Our result is robust to the case where $\tau_A \neq \tau_B$, as long as the gap $\tau_A - \tau_B$ is not too large. A formal proof is available upon request.

disagreement. While a full mechanism design is beyond the scope of this paper, our model suggests that platforms should aim to elicit private information about the true *intent* behind negative engagement, rather than simply suppressing tools like the "thumbs down" or dislike button, which may redirect backlash to the comment section (Kim et al., 2024).

Finally, our results challenge the common assumption that content-neutral moderation is ideologically neutral in its effects. Even when moderation targets only toxicity, the resulting shifts in visibility and engagement can systematically disadvantage ideological minorities. This concern is amplified in community-moderated platforms like Reddit, where moderation is decentralized and subreddit norms vary widely. For instance, Rajadesingan et al. (2021) show that politically oriented subreddits often apply rules asymmetrically, with content from ideological out-groups more likely to be flagged or removed. Additionally, users may strategically report ideologically opposing content to suppress dissent, a phenomenon often observed during polarized events such as elections or protests. Therefore, platform design must go beyond neutrality in policy and audit moderation decisions for fairness, not just accuracy.

4 Extension: Toxic Creators Valuing Negative Engagement

In the main model, we assume negative engagement discourages content creation by setting $\omega < 0$. This assumption aligns with many platforms' efforts to reduce the salience of negative feedback, such as YouTube's 2021 decision to make dislike counts private. This reflects concerns that visible disapproval deters participation. However, some studies suggest that some users—particularly toxic ones—may actually be motivated by negative responses. For instance, Buckels et al. (2014) find that certain toxic individuals derive pleasure from eliciting anger or outrage. Similarly, Cheng et al. (2014) document that negative votes ("dislikes") can provoke increased posting by low-quality contributors, sometimes by as much as 30%.

To capture this heterogeneity in motivation, we extend our model to allow negative engagement to encourage toxic users while continuing to discourage non-toxic users. Let $\omega(t)$ denote the weight placed on negative engagement by a type-t creator. For illustrative purposes, we set $\omega(NT) = \omega$ and $\omega(T) = -\omega$, with $\omega \in [-1, 0)$. The utility of content creator c of type (i, t) is then given by:

$$U^c(i,t) = S(t) \cdot P_s \cdot [L(i, NT) + \omega(t) \cdot DL(i, NT)] + \varepsilon_c \tag{7}$$

All other aspects of the model remain unchanged. The result is summarized in the following proposition.

Proposition 4. *When toxic users value negative engagement, the main results regarding non-toxic user's content creation* ($\frac{\partial P_c(A,NT)}{\partial \beta}$ *and* $\frac{\partial P_c(B,NT)}{\partial \beta}$)*, total positive engagement* (TL)*, and welfare inequality* ($\frac{\partial \mathbb{E}[(U^r(A)-U^r(B)]}{\partial \beta}$) *remain qualitatively unchanged.*

This extension confirms the robustness of our main findings. The intuition is as follows. Toxic users' preference for negative engagement alters their own creation incentives directly, but only indirectly influences their engagement behavior through changes in content supply. Since the direct effect dominates, the strategic environment faced by non-toxic users remains largely unchanged. As a result, key outcomes, such as total positive engagement and welfare inequality, continue to move in the same direction as in the baseline model.

5 Discussions of Model Variation

In this section, we discuss several model variations and their implications that are not explicitly captured in the baseline analysis.

Endogenous Consumption To highlight the main mechanisms, our baseline model abstracts from user attrition by assuming that readers remain on the platform even when they experience negative utility. This assumption can be justified by positing a small baseline utility from platform usage that is independent of content engagement. However, allowing for endogenous exit based on utility would only strengthen our core results. Suppose readers leave the platform if their realized utility U^r falls below an exogenous threshold $\underline{u}$. Incorporating this mechanism reinforces two key mechanisms. First, because minority users encounter more ideologically misaligned content, they have lower expected utility and are disproportionately likely to exit. This reduces the minority audience base, diminishing positive engagement for non-toxic minority creators and further discouraging their participation — accelerating the spiral of silence. This aligns with empirical evidence of self-censorship and attrition among disadvantaged groups exposed to persistent hostility (Haimson et al., 2021). Second, as minority users churn, the ideological composition becomes increasingly imbalanced—δ rises. A larger δ amplifies the asymmetry between in-group validation and out-group hostility, making future rounds of moderation more likely to fall into the "polarized expression" region described in Proposition 2. The model thus predicts a path-dependent escalation: small initial imbalances can, via *differential* exit, evolve into entrenched ideological dominance.

Content Filtering Beyond content removal, many platforms rely on content filtering— limiting the visibility of content whose toxicity score exceeds certain thresholds without deleting posts (Beknazar-Yuzbashev et al., 2025). In principle, these tools can potentially help offset some of the unintended effects from moderation. For example, platforms could increase visibility for non-toxic posts from under-represented groups until the risk of backlash falls below the threshold identified in our model. We do not model these filtering decisions explicitly, as doing so would require strong assumptions about platform objectives and user exposure. Instead, our analysis offers a foundation for when and how such interventions are most needed. Given the similar implications of content filtering and direct content removal, our results also reveal a key limitation: content filtering cannot fully correct the externalities faced by ideological minorities without promoting "echo chambers" by filtering out even non-toxic but ideologically opposed content: a practice ironically criticized for undermining diversity.

Advertising Incentives We deliberately abstract from the advertiser side of the market to isolate the strategic effects of content moderation on creator behavior. While platforms ultimately seek to maximize profit, positive engagement remains a central and observable proxy. Modeling the full two-sided market would require strong assumptions about advertiser preferences and pricing mechanisms, which risk obscuring our core contribution: showing how moderation, even absent explicit profit motives, can generate unintended negative externalities on ideological minorities. By focusing on this common and empirically grounded benchmark, we remain agnostic to firm's goal (e.g., maximizing engagement, creator, or reader welfare) and our analysis highlights a first-order concern that likely persists in more complex objective functions.

6 Conclusion

This paper offers a strategic perspective on toxicity in online platforms, showing that toxic content creation is not merely a static behavior to be suppressed, but a dynamic, incentive-driven response to the platform's moderation environment. Importantly, we show that moderation is not ideologically neutral, even when it targets only toxic content, because visibility and engagement are unequally distributed across ideological groups. As a result, the same policy can silence, stimulate, or reshuffle voices depending on the platform's polarization landscape.

These findings carry both managerial and policy implications. For platforms that rely on user-generated content, such as Wattpad and YouTube, sustaining active creation requires more than

removing harmful content; it demands designing incentive-compatible systems that separate toxicity from ideological disagreement. To complement such mechanisms, platforms may also explore ways to reduce affective polarization itself. For instance, by elevating norm-setting users with low toxicity and high credibility across ideological lines, or offering prompts that encourage users to frame disagreement constructively. From a regulatory standpoint, our results underscore the need to consider the structural and behavioral roots of polarization when evaluating fairness in content governance. Future work could extend our framework by modeling the design of truthful flagging mechanisms that reduce such negative externalities.

Our study also offers a few testable empirical predictions. For instance, we predict that identical moderation policies may produce asymmetric effects on content creation depending on a platform's ideological composition and the degree of affective polarization. Holding content quality constant, minority-group creators are likely to experience more negative engagement and sharper declines in creation following intensified moderation. We hope future research will examine these predictions in field or experimental settings to guide evidence-based content governance.

References

Andres, R. and Slivko, O. (2021). Combating online hate speech: The impact of legislation on twitter. Technical report, ZEW Discussion Papers.

Beknazar-Yuzbashev, G., Jiménez Durán, R., McCrosky, J., and Stalinski, M. (2025). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN 4307346*.

Berger, J. and Milkman, K. L. (2012). What makes online content viral? *Journal of Marketing Research*, 49(2):192–205.

Bondi, T., Rafieian, O., and Yao, Y. (2025). Privacy and polarization: An inference-based framework. *Management Science*.

Broniatowski, D. A. et al. (2023). The efficacy of facebook's vaccine misinformation policies and architecture during the covid-19 pandemic. *Science Advances*, 9(39):eadh2132.

Buckels, E. E., Trapnell, P. D., and Paulhus, D. L. (2014). Trolls just want to have fun. *Personality and Individual Differences*, 67:97–102.

Carlson, C. R. and Rousselle, H. (2020). Report and repeat: Investigating facebook's hate speech removal process. *First Monday*.

Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. (2014). How community feedback shapes user behavior. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8, pages 41–50.

Cima, L., Tessa, B., Cresci, S., Trujillo, A., and Avvenuti, M. (2024). Investigating the heterogeneous effects of a massive content moderation intervention via difference-in-differences. *arXiv preprint arXiv:2411.04037*.

Cinelli, M., De Francisci Morales, G., Galeazzi, A., Quattrociocchi, W., and Starnini, M. (2021). The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118.

Coase, R. H. (1960). The problem of social cost. *Journal of Law and Economics*, 3(1):1–44.

Dave, P. (2020). Social media giants warn of ai content moderation errors, as employees sent home. In *World Economic Forum. https://www. weforum. org/agenda/2020/03/social-media-giants-ai-moderation-errors-coronavirus/. Accessed December*, volume 27, page 2020.

Druckman, J. N. and Levendusky, M. S. (2019). What do we measure when we measure affective polarization? *Public Opinion Quarterly*, 83(1):114–122.

Godes, D., Mayzlin, D., Camara, O., Chung, D., Hydock, C., Kotchmar, R., Lim, C., Moshary, S., Paharia, N., Wernerfelt, N., et al. (2019). Politics, persuasion and choice. *Available at SSRN 3479876*.

Grossman, S. J. and Stiglitz, J. E. (1980). On the impossibility of informationally efficient markets. *The American Economic Review*, 70(3):393–408.

Haimson, O. L., Semrau, M., Matias, N., and Vitak, J. (2021). Disproportionate removals and differing content-moderation experiences for conservative, transgender, and black social media users. *Proceedings of the ACM on Human–Computer Interaction (CSCW)*, 5(CSCW2):Article 466.

Hickey, D., Schmitz, M., Fessler, D., Smaldino, P. E., Muric, G., and Burghardt, K. (2023). Auditing

elon musk's impact on hate speech and bots. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 17, pages 1133–1137.

Hodas, N. O. and Lerman, K. (2012). How visibility and divided attention constrain social contagion. In *2012 International Conference on Privacy, Security, Risk and Trust and 2012 International Confernece on Social Computing*, pages 249–257. IEEE.

Homroy, S. and Gangopadhyay, S. (2023). Political polarization and corporate political advocacy. *Available at SSRN 4742753*.

Huang, J. T., Choi, J., and Wan, Y. (2024). Politically biased moderation drives echo chamber formation: An analysis of user-driven content removals on reddit. *Available at SSRN*.

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

Jhaver, S., Appling, D. S., Gilbert, E., and Bruckman, A. (2019). " did you suspect the post would be removed?" understanding user reactions to content removals on reddit. *Proceedings of the ACM on Human-computer Interaction*, 3(CSCW):1–33.

Jiménez Durán, R. (2021). The economics of content moderation: Evidence from hate speech on twitter. *Available at SSRN 4044098*.

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

Kim, H., Lu, D., Ma, X., and Tafti, A. (2024). The impact of youtube's hiding dislike count on viewer and creator engagement. SSRN Working Paper.

Kozyreva, A., Herzog, S. M., Lewandowsky, S., Hertwig, R., Lorenz-Spreen, P., Leiser, M., and Reifler, J. (2023). Resolving content moderation dilemmas between free speech and harmful misinformation. *Proceedings of the National Academy of Sciences*, 120(7):e2210666120.

Lada, A., Wang, M., and Yan, T. (2021). How does news feed predict what you want to see? Meta Newsroom blog, accessed 30 June 2025.

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

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

Liaukonytė, J., Tuchman, A., and Zhu, X. (2023). Frontiers: Spilling the beans on political consumerism: Do social media boycotts and buycotts translate to real sales impact? *Marketing Science*, 42(1):11–25.

Liu, Y., Yildirim, P., and Zhang, Z. J. (2022). Implications of revenue models and technology for content moderation strategies. *Marketing Science*, 41(4):831–847.

Moorthy, K. S. (1985). Using game theory to model competition. *Journal of Marketing Research*, 22(3):262–282.

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

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

Noelle-Neumann, E. (1974). The spiral of silence a theory of public opinion. *Journal of Communication*, 24(2):43–51.

Our Mental Health (2025). Reddit downvote impact: User engagement & community dynamics. Accessed: 2025-03-16.

Pei, A. and Mayzlin, D. (2024). Do curation algorithms amplify the effect of trolls on users? Technical report, Working Paper.

Rajadesingan, A., Budak, C., and Resnick, P. (2021). Political discussion is abundant in non-political subreddits (and less toxic). In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 15, pages 525–536.

Schad, C. (2023). How the bud light boycott started—and why it's still going. NBC News, June 29, 2023.

Stsiampkouskaya, K., Joinson, A., and Piwek, L. (2023). To like or not to like? an experimental study on relational closeness, social grooming, reciprocity, and emotions in social media liking. *Journal of Computer-Mediated Communication*, 28(2):zmac036.

Thomas, D. R. and Wahedi, L. A. (2023). Disrupting hate: The effect of deplatforming hate organizations on their online audience. *Proceedings of the National Academy of Sciences*, 120(24):e2214080120.

Toubia, O. and Stephen, A. T. (2013). Intrinsic vs. image-related utility in social media: Why do people contribute content to twitter? *Marketing Science*, 32(3):368–392.

Twitter (2021). Introducing birdwatch, a community-based approach to misinformation. Accessed June 30, 2025.

Wuestenenk, N., van Tubergen, F., and Stark, T. H. (2025). The influence of group membership on online expressions and polarization on a discussion platform: An experimental study. *New Media & Society*, 27(1):225–245.

A Appendix

A.1 Equilibrium Characterization

There exists a unique solution to the system of equations outlined in (6), for each creator type, $P_c^*(i,t)$, given below:

$$\begin{aligned}
P_c^*(A,NT) &= \frac{-(\beta-1)x((3\beta-2)\omega+2)-2\omega+2\alpha(\omega-1)\left(2\delta^2+\delta-1\right)-12}{M_1} \\
&+ \frac{\alpha x(\omega-1)(\beta(\beta+\delta(\beta(4\tau_A-2)-8\tau_A+6)-3)+2)}{M_1} \\
P_c^*(B,NT) &= \frac{-(\beta-1)x((3\beta-2)\omega+2)-2\omega+2\alpha(\omega-1)(\delta-1)(2\delta+1)-12}{M_1} \\
&+ \frac{\alpha x(\omega-1)(\beta(\beta+\delta(\beta(4\tau_A-2)-8\tau_A+2)-3)+2)}{M_1} \\
P_c^*(A,T) &= \frac{-\omega((\alpha-3)\beta+4)+\alpha\beta+4\alpha(\omega-1)\delta^2+2\alpha\delta(\beta(-\omega)+\beta+\omega-1)-10}{M_1} \\
&+ \frac{x(3\beta\omega-2\omega+\alpha(\omega-1)(\beta(\delta(4\beta(\tau_A-1)-8\tau_A+6)-1)+2)+2)}{M_1} \\
P_c^*(B,T) &= \frac{-\omega((\alpha-3)\beta+4)+\alpha\beta+4\alpha(\omega-1)\delta^2+2\alpha(\beta-1)(\omega-1)\delta-10}{M_1} \\
&+ \frac{x(3\beta\omega-2\omega+\alpha(\omega-1)(\beta(4(\beta-2)\tau_A\delta+2\delta-1)+2)+2)}{M_1}
\end{aligned} \tag{A-1}$$

where

$$\begin{aligned}
M_1 = 2\Big(&x\big((-3(\beta-2)\beta-2)\,\omega+\alpha(\omega-1)\big((\beta-2)\beta\,(4\tau_A\delta-2\delta+1)+2\big)+2\big) \\
&-\,\omega+\alpha(\omega-1)\,(4\delta^2-1)-10\Big).
\end{aligned}$$

By substituting the creation probability derived above, we obtain the equilibrium expression for the visibility probability, P_s^*, as follows:

$$P_s^* = -\frac{4(\beta x+1)}{M_2} \tag{A-2}$$

where $M_2 = x((-3(\beta-2)\beta-2)\omega+\alpha(\omega-1)((\beta-2)\beta(4\tau_A\delta-2\delta+1)+2)+2)-\omega+\alpha(\omega-1)\left(4\delta^2-1\right)-10$. Rather than presenting the explicit expressions for total positive engagement (TL) and welfare inequality ($\mathbb{E}\left[U^r\left(A\right)-U^r\left(B\right)\right]$), we outline below the procedure used to compute these two metrics.

$$TL = \sum_{(i',t')\in\{A,B\}\times\{NT,T\}} Pr(i',t')\cdot\left[\left(\frac{1}{2}+\delta\right)\cdot P_l(A,i't')+\left(\frac{1}{2}-\delta\right)\cdot P_l(B,i't')\right] \tag{A-3}$$

and

$$\mathbb{E}\left[U^r\left(A\right)-U^r\left(B\right)\right]=\sum_{(i',t')\in\{A,B\}\times\{NT,T\}}Pr(i',t')\cdot\left[U^r\left(A,i't'\right)-U^r(B,i't')\right] \tag{A-4}$$

Note that, $Pr(i,t)=\frac{\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}{\sum_{i\in\{A,B\},t\in\{T,NT\}}\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}$ represents the actual probability that content of type (i,t) is viewed by a reader, given the equilibrium level of content creation. This differs from the linear approximation of visibility from the creator's perspective, which is employed to ensure analytical tractability in the model.

A.2 Proofs

We define the following notation used throughout the proofs. Let

$$v_T\equiv\sum_{i\in\{A,B\}}\lambda(i,T)\cdot S(T)\cdot P_c(i,T)$$
$$v_{NT}\equiv\sum_{i\in\{A,B\}}\lambda(i,NT)\cdot S\left(NT\right)\cdot P_c(i,NT)$$

denote, respectively, the total volume of toxic and non-toxic posts that survive moderation. Similarly, let

$$C(T)=\sum_{i\in\{A,B\}}\lambda(i,T)\cdot[L(i,T)+\omega DL(i,T)]$$
$$C\left(NT\right)=\sum_{i\in\{A,B\}}\lambda(i,NT)\cdot[L(i,NT)+\omega DL(i,NT)]$$

represent the total contribution to external engagement, conditional on exposure, from toxic and non-toxic users, respectively.

Lemma A.1. *Equilibrium visibility increases with the intensity of content moderation, i.e.,* $\frac{dP_s^*}{d\beta}>0$. *However, the expected exposure for toxic content decreases with the intensity of content moderation, i.e.,* $\frac{d[(1-\beta)P_s^*]}{d\beta}<0$.

Proof of Lemma A.1 The proof of this lemma proceeds in two steps.

Step 1. Show $\frac{dP_s^*}{d\beta}>0$.

According to the definition of visibility P_s^* and equation (6), the equilibrium content creation probability $P_c^*(i,t)$ is a function of equilibrium visibility P_s^*. Hence we can view the equilibrium visibility (P_s) as a fixed point.

Recall that, $P_s = 1 - \sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t)\cdot S(t)\cdot P_c(i,t)$.

Thus, we define $G(P_s^*,\beta) \equiv P_s^* - 1 + \sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t)\cdot S(t)\cdot P_c(i,t) = 0$. Based on the Implicit Function Theorem (IFT), we have

$$\frac{\partial P_s^*}{\partial \beta} = -\frac{\partial G/\partial \beta}{\partial G/\partial P_s^*}.$$

First, we can show that

$$\begin{aligned}
\frac{\partial G}{\partial \beta} &= \frac{\partial}{\partial \beta}\left[\sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t)\cdot S(t)\cdot P_c(i,t)\right] \\
&= \sum_{i\in\{A,B\}} \lambda(i,T)\cdot P_c(i,T)\cdot \frac{\partial S\,(T)}{\partial \beta} \\
&= -\sum_{i\in\{A,B\}} \lambda(i,T)\cdot P_c(i,T) \\
&< 0.
\end{aligned}$$

Then, we have

$$\begin{aligned}
\frac{\partial G}{\partial P_s^*} &= 1 + \frac{\partial}{\partial P_s^*}\left[\sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t)\cdot S(t)\cdot P_c(i,t)\right] \\
&= 1 + \sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t)\cdot S(t)\cdot \frac{\partial P_c(i,t)}{\partial P_s^*} \\
&= \frac{1}{4}\left(4\alpha\delta^2(1-\omega) - \alpha(1-\omega) + \alpha(1-\beta)x(1-\omega) + \alpha x(1-\omega) - 3\beta x\omega - 2x(1-\omega) + \omega + 6\right).
\end{aligned}$$

We can show that $\frac{\partial G}{\partial P_s^*}$ is a monotonic function of ω. This is because

$$\frac{\partial^2 G}{\partial P_s^* \partial \omega} = \frac{1}{4}\left(1 - 4\alpha\delta^2 + \alpha + x(2-\alpha(2-\beta))\right) > 0.$$

It then follows that

$$\begin{aligned}
\frac{\partial G}{\partial P_s^*} &\geq \frac{\partial G}{\partial P_s^*}|_{\omega=-1} \\
&= \frac{1}{4}\left(\alpha\left(8\delta^2 - 2\right) + x(\alpha x(4-\beta) - 4) + 5\right) \\
&\geq \frac{1}{4}(-2 - 4x + 5) \\
&> 0.
\end{aligned}$$

The second last inequality follows from $\delta \in [0, \frac{1}{2}], \alpha \in [0,1], \beta \in [0,1]$ and the last inequality follows $x \in (0, \frac{3}{4})$. As a result, we have $\frac{\partial P_s^*}{\partial \beta} = -\frac{\partial G/\partial \beta}{\partial G/\partial P_s^*} > 0$.

Step 2. Show $\frac{d[(1-\beta)P_s^*]}{d\beta} < 0$.

Define $F(\beta, P_s^*) = (1-\beta)P_s^*$. Taking derivatives of $F(\cdot)$ w.r.t. β, we have

$$\frac{\partial F}{\partial \beta} = -P_s^* + (1-\beta)\frac{\partial P_s^*}{\partial \beta} \tag{A-5}$$

Recall that $G(P_s^*, \beta) = P_s^* - 1 + \sum_{i\in\{A,B\},t\in\{T,NT\}} \lambda(i,t) \cdot S(t) \cdot P_c(i,t) = 0$. Then we can rewrite $G(\cdot)$ as follows

$$G(P_s^*, \beta) = P_s^* - 1 + v_T + v_{NT} = 0,$$

where v_T and v_{NT} refer to, respectively, the total volume of toxic and non-toxic posts that survive moderation. Taking derivatives of G w.r.t. β, we obtain

$$\frac{\partial G}{\partial \beta} = \frac{\partial P_s^*}{\partial \beta} + \frac{\partial v_T}{\partial \beta} + \frac{\partial v_{NT}}{\partial \beta} = 0 \tag{A-6}$$

By construction, we can further re-write $\frac{\partial v_T}{\partial \beta}$ and $\frac{\partial v_{NT}}{\partial \beta}$ as a function of F and P_s^* as follows:

$$\frac{\partial v_T}{\partial \beta} = C(T) \cdot \frac{\partial F(\beta, P_s^*)}{\partial \beta} \tag{A-7}$$

$$\frac{\partial v_{NT}}{\partial \beta} = C(NT) \cdot \frac{\partial P_s^*}{\partial \beta}, \tag{A-8}$$

where $C(T) = \sum_{i\in\{A,B\}} \lambda(i,T)\cdot[L(i,T)+\omega DL(i,T)]$ and $C\left(NT\right) = \sum_{i\in\{A,B\}} \lambda(i,NT)\cdot[L(i,NT)+\omega DL(i,NT)]$, neither of which depends on β. Substituting them back into (A-6), we can solve for $\frac{\partial P_s^*}{\partial \beta}$ as a function of $\frac{\partial F}{\partial \beta}$:

$$\frac{\partial P_s^*}{\partial \beta} = \frac{-C(T)}{C\left(NT\right)+1} \cdot \frac{\partial F}{\partial \beta}. \tag{A-9}$$

Substituting it back to equation (A-5), we can finally derive the expressions of $\frac{\partial F}{\partial \beta}$ as follows

$$\frac{\partial F}{\partial \beta} = \frac{-P_s^*\left(C\left(NT\right)+1\right)}{C\left(NT\right)+1+\left(1-\beta\right)C(T)}. \tag{A-10}$$

It is easy to see that the denominator $C\left(NT\right)+1+\left(1-\beta\right)C(T) = \frac{\partial G}{\partial P_s^*} > 0$, which we have shown above. Hence the sign of $\partial F/\partial\beta$ depends on the sign of the numerator, $-P_s^*\left(C\left(NT\right)+1\right)$. With some simple algebra, we can show that:

$$\begin{aligned} C\left(NT\right)+1 =& \frac{1}{4}\left(\alpha\left(4\delta^2-1\right)(1-\omega)+\alpha x(1-\omega)+(1-x)\omega-2x+6\right) \\ \geq& \frac{1}{4}\left(-2-1-2x+6\right) \\ >& 0. \end{aligned}$$

The second last inequality follows from $\delta \in [0, \frac{1}{2}], \alpha \in [0,1]$ and $\omega \in [-1, 0)$ and the last inequality follows $x \in (0, \frac{3}{4})$. Hence, we have $\frac{\partial F}{\partial \beta} = \frac{d[(1-\beta)P_s^*]}{d\beta} < 0$. □

Proof of Result 1 When $\alpha = 0$, readers' content evaluation is unaffected by ideological affiliation and depends solely on the toxicity of the content. In other words, from the creator's perspective, expected engagement is determined entirely by the toxicity level of their content, regardless of their ideological group. As a result, the effect of moderation on content creation is symmetric across ideological groups. That is, we have:

$$P_c^*(A,t) = P_c^*(B,t), \forall t \in \{T, NT\}, \text{ and } \mathbb{E}[U^r(A)] = \mathbb{E}[U^r(B)].$$

We prove the four claims outlined in Result 1 below.

(i) $\frac{dP_c^*(i,NT)}{d\beta} > 0$

From equation (6), the derivative of $P_c^*(i, NT)$ with respect to β is:

$$\frac{\partial P_c^*(i,NT)}{\partial \beta} = \frac{1}{2}\frac{\partial P_s^*}{\partial \beta} \cdot \left(L(i,NT) + \omega \cdot DL(i,NT)\right).$$

Since $L(i,NT)+\omega \cdot DL(i,NT) = \frac{1}{4}(2+\omega) > 0$ for $\omega \in [-1,0)$, and Lemma A.1 establishes $\frac{\partial P_s^*}{\partial \beta} > 0$, the derivative is strictly positive.

(ii) $\frac{dP_c^*(i,NT)}{d\beta} > 0$ and $\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial \beta} < 0$

First, recall that

$$P_c^*(i,T) = \frac{1}{2} + \frac{1}{2} \cdot S(T) \cdot P_s^* \cdot \left(L(i,T) + \omega \cdot DL(i,T)\right).$$

Since $L(i,T) + \omega \cdot DL(i,T) = \frac{3\omega}{4} < 0$ for $\omega \in [-1,0)$, we have:

$$\frac{\partial P_c^*(i,T)}{\partial \beta} = \frac{1}{2}\frac{\partial[(1-\beta)P_s^*]}{\partial \beta} \cdot \frac{3\omega}{4} > 0.$$

The inequality directly follows Lemma Lemma A.1 as it establishes $\frac{\partial[(1-\beta)P_s^*]}{\partial \beta} < 0$.

Now take the derivative of the *survived* toxic content created in equilibrium:

$$\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial \beta} = -P_c^*(i,T) + (1-\beta) \cdot \frac{\partial P_c^*(i,T)}{\partial \beta}.$$

Using the expressions from Lemma A.1, we can rewrite the above as follows

$$\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial \beta} = -\frac{1}{2}\left(1 + P_c^*(i,T) \cdot \left(1 + \underbrace{\frac{1 + C\left(NT\right)}{C\left(NT\right) + 1 + (1-\beta)\,C(T)}}_{>0}\right)\right) < 0,$$

The derivative is strictly negative because $P_c^*(i,T) > 0$.

(iii) $\frac{\partial \mathbb{E}U^r(i)}{\partial \beta} > 0$

Under $\alpha = 0$, readers' utility depends only on content toxicity. For a representative user in ideological group A:

$$\mathbb{E}[U^r(A)] = Pr(A,T)\cdot(-1) + Pr(B,T)\cdot(-1).$$

Hence,

$$\begin{aligned}
\frac{\partial[\mathbb{E}U^r(A)]}{\partial\beta} &= -\left(\frac{\partial Pr(A,T)}{\partial\beta} + \frac{\partial P_r(B,T)}{\partial\beta}\right) \\
&= -\frac{x}{2(1-P_s^*)^2}\left(2(2P_s^*-1)P_c(A,T) + (1+(1-\beta)(L(A,T)+\omega DL(A,T))\cdot\frac{\partial[(1-\beta)P_s^*]}{\partial\beta}\right) \\
&= -\frac{xP_c(A,T)}{(1-P_s^*)^2}\left(2P_s^* - 1 - \frac{C\,(NT)+1}{C\,(NT)+1+(1-\beta)\,C(T)}\right) \\
&\geq -\frac{xP_c(A,T)}{(1-P_s^*)^2}\,(2P_s^*-1-1) \\
&= \frac{2xP_c(A,T)}{1-P_s^*} \\
&> 0
\end{aligned}$$

The third last inequality follows from Proof for Lemma A.1, where we can derive that

$$\begin{aligned}
\frac{C\,(NT)+1}{C\,(NT)+1+(1-\beta)\,C(T)} &= \frac{-x(\omega+2)+\omega+6}{2x(\omega-1)+\omega+6} \\
&\geq \frac{-x(\omega+2)+\omega+6}{2x(\omega-1)+\omega+6}\Big|_{\omega=0,x=0} \\
&= 1.
\end{aligned}$$

(iv) $\frac{\partial TL}{\partial\beta} > 0$

With $\alpha = 0$, positive engagement ("likes") only occur on non-toxic content, that is,

$$P_l(i,iNT) = P_l(i,i'NT) = \frac{1}{2}, \quad P_l(i,iT) = P_l(i,i'T) = 0, \forall i \neq i'.$$

Thus, total positive engagement is:

$$TL = \frac{1}{2}\left[Pr(A,NT) + Pr(B,NT)\right].$$

Hence, we have

$$\frac{\partial TL}{\partial\beta} = \frac{1}{2}\left(\frac{\partial Pr(A,NT)}{\partial\beta} + \frac{\partial Pr(B,NT)}{\partial\beta}\right).$$

Recall that

$$Pr(i,NT) = \frac{\lambda(i,NT)P_c(i,NT)}{1-P_s^*}, \forall i \in \{A,B\}.$$

and thus

$$\frac{\partial Pr(i,NT)}{\partial \beta} = \frac{1}{2}\frac{\lambda(i,NT)}{(1-P_s^*)^2}\frac{\partial P_s^*}{\partial \beta}(1 + L(i,NT) + \omega DL(i,NT)) \tag{A-11}$$

$$= \frac{1}{2}\frac{\lambda(i,NT)}{(1-P_s^*)^2}\frac{\partial P_s^*}{\partial \beta} \cdot \left(\frac{6+\omega}{4}\right) > 0 \tag{A-12}$$

The inequality follows Lemma A.1 which establishes $\frac{\partial P_s^*}{\partial \beta} > 0$. Hence, it directly follows that $\frac{\partial TL}{\partial \beta} > 0$. □

Proof of Proposition 1 With $\delta = 0$, the two ideological groups are symmetric. So, the effect of moderation on content creation should be the same across the two groups. Hence, $P_c^*(A,NT) = P_c^*(B,NT)$ and $P_c^*(A,T) = P_c^*(B,T)$. Taking derivative of $P_c^*(i,NT)$ w.r.t. β, we obtain

$$\frac{dP_c^*(i,NT)}{d\beta} = \frac{1}{2}\frac{dP_s^*}{d\beta} \cdot (L(i,NT) + \omega \cdot DL(i,NT))$$

$$= \frac{1}{2}\frac{dP_s^*}{d\beta} \cdot \left(\frac{\omega + \alpha(\omega-1) + 2}{4}\right).$$

According to Lemma A.1, we know that $\frac{dP_s^*}{d\beta} > 0$. Therefore, it directly follows that $\frac{dP_c^*(i,NT)}{d\beta} < 0$ iff $\alpha > \bar{\alpha}_1(\omega) \equiv \frac{\omega+2}{1-\omega}$. Otherwise, we have $\frac{dP_c^*(i,NT)}{d\beta} \geq 0$. It is easy to verify that $\bar{\alpha}_1(\omega) \leq 1$ if $\omega \in [-1, -\frac{1}{2}]$.

Next, taking derivative of $P_c^*(i,T)$ w.r.t. β, we obtain

$$\frac{dP_c^*(i,T)}{d\beta} = \frac{1}{2}\frac{d[(1-\beta)P_s^*]}{d\beta} \cdot (L(i,T) + \omega \cdot DL(i,T))$$

$$= \frac{1}{2}\frac{d[(1-\beta)P_s^*]}{d\beta} \cdot \left(\frac{\alpha(1-\omega) + 3\omega}{4}\right).$$

Lemma A.1 shows $\frac{d[(1-\beta)P_s^*]}{d\beta} < 0$. Therefore, it directly follows that $\frac{dP_c^*(i,NT)}{d\beta} > 0$ iff $\alpha < \bar{\alpha}_2(\omega) \equiv \frac{-3\omega}{1-\omega}$. Otherwise, we have $\frac{dP_c^*(i,NT)}{d\beta} \leq 0$. Moreover, it is easy to verify that $\bar{\alpha}_2(\omega) \leq 1$ if $\omega \in [-\frac{1}{2}, 0)$. □

Proof of Proposition 2 Taking derivative of $P_c^*(A,NT)$ w.r.t. β, we obtain

$$\frac{\partial P_c^*(A,NT)}{\partial \beta} = \frac{1}{2}\frac{\partial P_s^*}{\partial \beta} \cdot (L(A,NT) + \omega \cdot DL(A,NT))$$

$$= \frac{1}{2} \cdot \frac{\partial P_s^*}{\partial \beta} \cdot [\frac{\omega + \alpha(\omega-1)(1-2\delta) + 2}{4}]$$

which is negative if and only if $\alpha > \bar{\alpha} \equiv \frac{\omega+2}{(1-\omega)(1-2\delta)}$.

Taking derivative of $P_c^*(B, NT)$, we have

$$\begin{aligned}\frac{\partial P_c^*(B,NT)}{\partial \beta} =& \frac{1}{2}\frac{\partial P_s^*}{\partial \beta}\cdot(L(B,NT)+\omega\cdot DL(B,NT)) \\ =& \frac{1}{2}\cdot\frac{\partial P_s^*}{\partial \beta}\cdot[\frac{\omega+\alpha(\omega-1)(2\delta+1)+2}{4}]\end{aligned}$$

which is negative if and only if $\alpha > \underline{\alpha} \equiv \frac{\omega+2}{(2\delta+1)(1-\omega)}$.

According to Lemma A.1, we know that $\frac{dP_s^*}{d\beta} > 0$. Thus, it is easy to verify that $\bar{\alpha} > \underline{\alpha}$ always holds. Comparing the thresholds that determine the respective sign of $\frac{dP_c^*(A,NT)}{d\beta}$ and $\frac{dP_c^*(B,NT)}{d\beta}$ naturally gives us the three regions characterized in Proposition 2. □

Proof of Proposition 3 We prove the two claims outlined in Proposition 3 below.

(i) $\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta} > 0$

Based on the definition of welfare inequality in equation (A-4), we can rewrite it as follows:

$$\begin{aligned}\mathbb{E}[U^r(A)-U^r(B)] =& \, Pr(A,NT)[U^r(A,ANT)-U^r(B,ANT)] \\ &+ Pr(B,NT)[U^r(A,BNT)-U^r(B,BNT)] \\ &+ Pr(A,T)[U^r(A,AT)-U^r(B,AT)] \\ &+ Pr(B,T)[U^r(A,BT)-U^r(B,BT)],\end{aligned}$$

where $U^r(A,ANT) - U^r(B,ANT) = U^r(A,AT) - U^r(B,AT) = \alpha > 0$ and $U^r(A,BNT) - U^r(B,BNT) = U^r(A,BT) - U^r(B,BT) = -\alpha < 0$.

Thus, the effect of moderation on welfare inequality is given below,

$$\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta} = \alpha\left(\frac{\partial Pr(A,NT)}{\partial \beta} - \frac{\partial Pr(B,NT)}{\partial \beta} + \frac{\partial Pr(A,T)}{\partial \beta} - \frac{\partial Pr(B,T)}{\partial \beta}\right). \quad \text{(A-13)}$$

Note that the above equation simply represents the sum of marginal change in "exposure inequality", from the perspective of a reader, between the two ideological groups, aggregated over both toxic and non-toxic content in equilibrium.

Let

$$\begin{aligned}\Delta Pc(NT) &\equiv \lambda(A,NT)P_c^*(A,NT) - \lambda(B,NT)P_c^*(B,NT) \\ \Delta Pc(T) &\equiv S(T)\left[\lambda(A,T)P_c^*(A,T) - \lambda(B,T)P_c^*(B,T)\right]\end{aligned}$$

and recall that

$$Pr(i,t) = \frac{\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}{\sum_{i\in\{A,B\},t\in\{T,NT\}}\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}.$$

According to equation (A-13), we have

$$\begin{aligned}\frac{\partial \mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta} &= \frac{\partial[\frac{\Delta Pc(NT)+\Delta Pc(T)}{1-P_s^*}]}{\partial \beta} \\ &= \frac{\partial\left(\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right)}{\partial \beta} \cdot \frac{1}{1-P_s^*} - \frac{\left[\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right]}{(1-P_s^*)^2} \cdot \frac{\partial\left[1-P_s^*\right]}{\partial \beta}\end{aligned}$$

That is, proving $\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta} > 0$ is equivalent to proving the following inequality

$$\frac{\partial\left[\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right]}{\partial \beta} > \frac{\partial\left[1-P_s^*\right]}{\partial \beta} \cdot \frac{1}{1-P_s^*} \cdot \left[\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right]. \tag{A-14}$$

Note that we have $\frac{\partial[1-P_s^*]}{\partial \beta} < 0$ as established in Lemma A.1. First, we can expand $\Delta P_c(NT) + \Delta P_c(T)$ as follows:

$$\begin{aligned}\Delta P_c(NT) + \Delta P_c(T) = &\frac{P_s^*}{2} \cdot \lambda(A, NT)\left(1 + L(A, NT) + \omega \cdot DL(A, NT)\right) \\ &- \frac{P_s^*}{2}\lambda(B, NT)\left(1 + L(B, NT) + \omega \cdot DL(B, NT)\right) \\ &+ \frac{P_s^*}{2} \cdot \lambda(A, T)\left(1 + (1-\beta)(L(A, T) + \omega \cdot DL(A, T))\right) \\ &- \frac{P_s^*}{2} \cdot \lambda(B, T)\left(1 + (1-\beta)(L(B, T) + \omega \cdot DL(B, T))\right) \\ &= \frac{P_s^*}{2} \cdot N_1\end{aligned}$$

where $N_1 = \frac{1}{2}\delta(\alpha\beta x(\omega - 1) + \omega + 6) \geq 0$. Hence we can re-arrange inequality (A-14) as follows:

$$\begin{aligned}&\frac{\partial\left[\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right]}{\partial \beta} - \frac{\partial\left[1-P_s^*\right]}{\partial \beta}\frac{1}{1-P_s^*} \cdot \left[\Delta Pc\left(NT\right) + \Delta Pc\left(T\right)\right] \\ =&\frac{P_s^*}{2} \cdot \frac{\partial N_1}{\partial \beta} + \frac{1}{2} \cdot \frac{\partial P_s^*}{\partial \beta} \cdot N_1 + \frac{1}{2} \cdot \frac{\partial P_s^*}{\partial \beta}\frac{P_s^*}{1-P_s^*} \cdot N_1 \\ =&\frac{1}{2} \cdot \left[\frac{\partial N_1}{\partial \beta}P_s^* + \frac{\partial P_s^*}{\partial \beta} \cdot N_1 * (1 + \frac{P_s^*}{1-P_s^*})\right] \\ =&\frac{P_s^*}{2} \cdot \left[\frac{\partial N_1}{\partial \beta} + \frac{\partial P_s^*}{\partial \beta} \cdot N_1 * \frac{1}{P_s^*(1-P_s^*)})\right] \\ >&\frac{P_s^*}{2} \cdot \left[\frac{\partial N_1}{\partial \beta} + 4 \cdot \frac{\partial P_s^*}{\partial \beta} \cdot N_1\right] \\ >&\frac{P_s^*}{2} \cdot \left[\frac{\partial N_1}{\partial \beta} + 4 \cdot \frac{\partial P_s^*}{\partial \beta} \cdot N_1\right]|_{\delta=0} \\ =&0\end{aligned}$$

The second last inequality follows that the term in the bracket monotonically increases in δ.

(ii) $\frac{\partial TL}{\partial \beta} > 0$

Given the respective share of users, we can rewrite total positive engagement (TL) as follows:

$$\begin{aligned}TL =& \frac{1}{4}[Pr(A,NT)(\alpha(2\delta-1)+2)+Pr(A,T)\alpha(2\delta+1)]\\ &+\frac{1}{4}[Pr(B,NT)(2-\alpha(2\delta+1))+Pr(B,T)(\alpha(1-2\delta)]\\ =& \frac{1}{4}[\alpha(1-2\delta)+2(1-\alpha)(Pr(A,NT)+Pr(B,NT))+4\alpha\delta(Pr(A,NT)+Pr(A,T))]\end{aligned}$$

Taking derivatives wrt β, we obtain:

$$\frac{\partial TL}{\partial \beta} = \frac{1}{2}(1-\alpha)\left(\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(B,NT)}{\partial \beta}\right)+\alpha\delta\left(\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(A,T)}{\partial \beta}\right).$$

It is straightforward to see that $\frac{\partial TL}{\partial \beta} > 0$ holds if both

$$\left(\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(B,NT)}{\partial \beta}\right) > 0 \text{ and } \left(\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(A,T)}{\partial \beta}\right) > 0.$$

We now show, in sequence, that both conditions are satisfied.

First, we can show that

$$\frac{\partial Pr(i,NT)}{\partial \beta} = \frac{\lambda(i,NT)\frac{\partial P_s^*}{\partial \beta}}{2(1-P_s^*)^2}\cdot[1+L(i,NT)+\omega DL(i,NT)]$$

where

$$\begin{aligned}1+L(A,NT)+\omega DL(A,NT) &= \frac{1}{4}(\omega+\alpha\delta(\omega-2(\omega-1)-1)+6)\\ &\geq \frac{1}{4}(-1-3\alpha+6)\\ &> 0\\ 1+L(B,NT)+\omega DL(B,NT) &= \frac{1}{4}(\omega+\alpha(\omega-1)(2\delta+1)+6)\\ &\geq \frac{1}{4}(-1-4\alpha+6)\\ &> 0\end{aligned}$$

Hence, we have $\frac{\partial Pr(i,NT)}{\partial \beta} > 0$ for $i \in [A,B]$. Then we must have $\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(B,NT)}{\partial \beta} > 0$.

Second, according to the above proof of welfare inequality, we know that the following is true

$$\frac{\partial Pr(A,NT)}{\partial \beta}-\frac{\partial Pr(B,NT)}{\partial \beta}+\frac{\partial Pr(A,T)}{\partial \beta}-\frac{\partial Pr(B,T)}{\partial \beta} > 0. \tag{A-15}$$

Moreover, given that these probabilities add up to 1, i.e., $\sum_{i\in\{A,B\},t\in\{T,NT\}} Pr(i,t) = 1$, we must have

$$\frac{\partial Pr(A,NT)}{\partial \beta}+\frac{\partial Pr(B,NT)}{\partial \beta}+\frac{\partial Pr(A,T)}{\partial \beta}+\frac{\partial Pr(B,T)}{\partial \beta} = 0 \tag{A-16}$$

Combining (A-15) and (A-16), we must have $\frac{\partial[Pr(A,NT)+Pr(A,T)]}{\partial \beta} > 0$. □

Proof of Proposition 4 The proof for content creation is identical to that of Proposition 2, since engagement from other users depends only on their group affiliations. The only difference is the expression for $\frac{\partial G}{\partial P_s^*}$, shown below.

$$\frac{\partial G}{\partial P_s^*} = \frac{1}{4}\left(x((3-\alpha)\beta\omega + \alpha(2-\beta) - 4\omega - 2) + \omega - \left(\alpha(1-\omega)\left(1-4\delta^2\right)\right) + 6\right).$$

We can show that $\frac{\partial G}{\partial P_s^*}$ is a monotonic increasing function of ω. Thus,

$$\begin{aligned}
\frac{\partial G}{\partial P_s^*} &\geq \frac{\partial G}{\partial P_s^*}|_{\omega=-1} \\
&= \frac{1}{4}\left(\alpha\left(8\delta^2-2\right) + x(2\alpha - 3\beta + 2) + 5\right) \\
&\geq \frac{1}{4}(-2-x+5) \\
&> 0.
\end{aligned}$$

The rest of the proof and expressions are identical.

The key insight is that allowing toxic users to derive utility from negative engagement affects only their content creation, not their *engagement* with others' posts. This behavioral shift changes the equilibrium quantity of toxic content, and thus the composition of visible content, but it does not reverse the directional effect of moderation. In particular, we still have:

$$\frac{\partial[(1-\beta)P_s^*]}{\partial\beta} < 0.$$

Hence, the direct reduction in exposure to toxic content caused by stricter moderation outweighs any indirect adjustments in the supply of toxic content.

Because both reader welfare and total positive engagement are determined from the consumption side—that is, they are mediated by post visibility—their comparative-static responses to β follow immediately from the same logic as documented in Proposition 3.

□