CES ifo
Working Papers

12008
2025

Original Version: July 2025
This Version: November 2025

Spiral of Silence: How Neutral Moderation Polarizes Content Creation

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:

How Neutral Moderation Polarizes Content Creation *

Ying Bao
University of Illinois Urbana-Champaign

Jessie Liu
Johns Hopkins University

This Version: November, 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, i.e., users' aversion to ideologically opposed content, interacts with moderation in unintended ways. We show that 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 creators from the ideological majority group sometimes benefit from this exposure, they do not internalize the negative spillovers, i.e., increased out-group animosity toward minority creators. This can discourage minority creation and polarize the content supply, ultimately leaving minority users in a more ideologically imbalanced environment: a mechanism reminiscent of the "spiral of silence." Thus, our model offers an alternative perspective to a common debate: what appears as bias in moderation needs not reflect bias in rules, but can instead emerge endogenously as self-censorship in equilibrium. We also extend the model to explore how content personalization interacts with moderation policies.

Keywords: Content Moderation, Polarization, Platform Governance, 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, the Marketing Theory Seminar, and seminars at the University of Washington and University of Illinois at Urbana-Champaign. We also thank Rafael Jiménez-Durán, David Godes, Michael Keane, Peter Landry, Alexei Makarin, Cameron Martel, Sridhar Moorthy, 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 (2021), 4 in 10 Americans have faced online harassment, and 71% support tighter platform rules on toxic content. This growing concern is reflected in emerging regulatory frameworks worldwide, including German Network Enforcement Act (NetzDG) and the European Union's Digital Services Act (DSA).[1] Existing research on content moderation largely focuses on the demand side: how users respond when speech is restricted. Prior work has shown that content removal can trigger backlash (Jhaver et al., 2019), increase 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 the *anticipation* of moderation shapes what content is produced 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 what content is *never created* at all.

A key motivation for our approach comes from a puzzling empirical pattern: opposite moderation policies can sometimes produce similar outcomes. When Reddit banned several toxic communities during "The Great Ban" in 2020, some former toxic users became more active, but most, especially the less-toxic ones, posted less or stopped altogether (Cima et al., 2024a,b). Conversely, when Twitter relaxed its rules in 2022, hate speech 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

[1] https://digital-strategy.ec.europa.eu/en/library/code-conduct-countering-illegal-hate-speech-online

of content, they are also strategic creators who weigh the risks and rewards of content creation.

Despite growing attention to polarization, most existing research conflate ideological division with toxicity, treating both as a single notion of "offensiveness." This theoretical conflation obscures the distinct incentives and consequences that arise from ideology-based versus toxicity-based preferences. In practice, however, algorithm-based moderation systems mainly targets the latter. For instance, widely used Google's Perspective API are designed to be ideologically neutral, trained on data across the political spectrum and calibrated to detect toxic language rather than viewpoint (Rieder and Skop, 2021). Yet, content that is non-toxic in language can still provoke severe backlash when it expresses ideological disagreement, depending on *who* is reading. For example, TikTok influencer Leo Skepi, with over 4 million followers, faced strong 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 formalizes this distinction along two aspects of user preferences: (i) a *vertical* preference capturing aversion to overtly toxic content, and (ii) a *horizontal* preference capturing aversion to ideologically opposed content. This framework allows us to identify the conditions under which ex ante "neutral" moderation, by reweighting anticipated engagement, can produce systematically non-neutral outcomes ex post.

In this paper, we develop a model in which users act as both consumers and (potential) creators of content, differing along two dimensions: their ideological position and their propensity to produce toxic content. Content creation depends on both intrinsic motivation and external engagement, whether positive or negative, from ideologically aligned (in-group) and opposing (out-group) audiences. A central feature of our framework is explicitly modeling *affective polarization*: the extent to which users respond emotionally to the ideological identity of content sources, rather than to the content's tone or language itself (Iyengar et al., 2019). Affective polarization gives rise to systematic *in-group favoritism* and *out-group animosity*: toxic posts often receive some tolerance within ideological groups, whereas even civil content can 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 negatively when it comes from ideological out-group members (Wuestenenk et al., 2025). Alongside this, our model captures how *ideological imbalance*, i.e., unequal group sizes within the user base, creates asymmetry in who receives more validation versus hostility. Together, these two forces give rise to what we term as *relational externality*: any

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

policy that alters exposure, such as moderation or personalization, reshapes the balance between in-group favoritism and out-group animosity that governs creators' incentives. In other words, a policy aimed at regulating one dimension of content (e.g., toxicity) can indirectly influence other aspects (e.g., ideology) of content creation by altering the social (relational) environment in which content is rewarded or penalized.

Our analysis offers two key insights. First, when both affective polarization and ideological imbalance are high, moderation amplifies in-group favoritism and out-group animosity through increasing the content reach of non-toxic posts from all ideological groups. Facing a larger in-group, this motivates the ideological majority creators to produce more content, whereas the ideological minority are discouraged from creating, as they expect an intensified animosity from a larger out-group. This dynamic reproduces a self-reinforcing "spiral of silence" (Noelle-Neumann, 1974): what appears as ideological bias in moderation outcomes can emerge *endogenously* in equilibrium, even when moderation rules are designed to be ideologically neutral. This insight also challenges the very notion of "neutrality" in policy design. Content-neutral moderation may not be outcome-neutral because it changes the relational environment that determines equilibrium content creation.

Second, moderation may improve average outcomes but redistributes welfare unevenly across consumers: 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. In other words, majority content gain higher reach without internalizing its negative spillover on the minority. These results have direct implications for policy frameworks such as the European Union's 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. Without accounting for the distinct roles of ideology and toxicity in user behavior, moderation policies may unintentionally reinforce polarization and marginalize civil under-represented groups.

We also extend the baseline model along four dimensions to test the policy relevance and robustness of our results. The first two extensions examine content personalization and the presence of ideologically neutral users, which are interventions often proposed as immediate remedies to the spiral of silence. We show that personalization protects minority creators from out-group animosity but narrows their reach, and by the same logic can also revive toxicity within ideological groups.

Likewise, while a large group of neutral users can diffuse animosity and promote creation, a small one may deepen silence among those very neutral users meant to bridge ideological divides. The final two extensions introduce alternative motivations for toxic users, allowing them to value either toxic consumption or negative engagement. Across both cases, our core insight holds: moderation continues to shape content creation through the same underlying relational externality.

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. 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 be abused and 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, 2023b), 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; Berman and Katona, 2020). Although most 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. One exception is Berman and Katona (2020), which highlights how algorithmic curation affects both the diversity and quality of content consumed. Our model complements this perspective by identifying a structural "market failure" that arises not from misaligned incentives between the platform and users, but from polarized content supply driven by asymmetric in-group versus out-group engagement. This externality persists regardless of *who* controls moderation or personalization.

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 expected reach and 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.

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). Affective polarization generalizes beyond the simple in-group/out-group dichotomy in political debates and can be viewed as a "structural property of social networks" (Lerman et al., 2024). Our model captures this network feature by linking users' creation and engagement incentives to the relational environment shaped by others' reactions. This distinction is particularly 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) 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 features 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 explore a few important extensions of the baseline model. Section 5 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. Here, we model toxicity as an exogenous and fixed consumer type, consistent with prior psychology literature that link online trolling to stable personality traits. For instance, Buckels et al. (2014) and Craker and March (2016) show that toxic online behavior is strongly associated with enduring dark traits such as sadism, suggesting that the *differential* propensity to engage in toxicity reflects dispositional rather than situational factors. 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 $\tau_A = 1$, all toxic users are from group A; when $\tau_A = 0$, all toxic users are from group B. Hereafter in the discussion, for clarity, we always refer to group A as the (ideological) majority group and B as the minority group.

	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. Table 1 summarizes the resulting population mass of each user type.

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 creator c of type (i',t') as:

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

where

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

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 opposing content, 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, in our baseline model, the utility specification implies 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 reflects the assumption that users evaluate others' content based on its ideology and tone, rather than their *own* posting behavior. In other words, we treat content toxicity as a vertical attribute in the baseline model: users who engage in toxic creation also experience disutility from consuming toxic content, particularly when it originates from the ideological out-group (Rabbani and Pusch, 2025). In Section 4.3, we relax this assumption of toxicity as a purely vertical attribute and allow toxic users to derive utility from toxicity.

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.

[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.

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

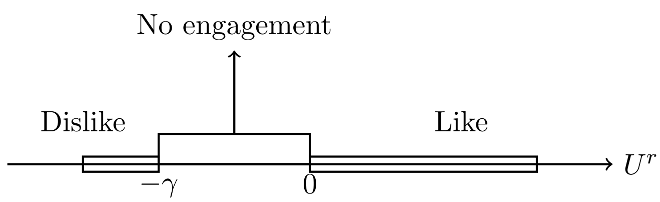


Figure 1: User Engagement Based on $U^r(\cdot)$

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.

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

$P_l(r,c)$		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 (i.e., "like") non-toxic content created by users with the same ideology. Meanwhile, users are more likely to negatively engage with (i.e., "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 (Toubia and Stephen, 2013). We formalize the utility of content creation for a user (creator) c of type (i,t) as follows:

$$U^c(i,t) = \underbrace{S(t)}_{\text{survival rate}} \cdot \underbrace{R(V)}_{\text{content reach}} \cdot \underbrace{(NE^{in}(i,t) + NE^{out}(i,t))}_{\text{in- and out-group engagement}} \underbrace{+\varepsilon_c}_{\text{intrinsic utility}}, \tag{2}$$

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 able to reach audiences. We model moderation as acting on posts rather than on reactions. This matches empirical practice: hate speech and toxicity are measured and reported primarily at the level of top-level posts, as in Hickey et al. (2023)'s analysis of toxic tweet spikes and Müller and Schwarz (2021, 2023a)'s measurement of hate tweets. Moreover, Beknazar-Yuzbashev et al. (2025) show that widely used moderation tools, such as Perspective-based filters, operate mainly through post-

level adjustments, whereas replies receive far less exposure and are not the primary target of automated moderation.[5] 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 $1 - S(t)$.

2. Content Reach ($R(V)$): Once live, content competes for user attention. We assume that each surviving post attains a uniform reach across the platform's readership. Section 4.1 relaxes this assumption by introducing content personalization. The platform's exposure mechanism is modeled through a linear "reach factor" ($R(\cdot)$) that decreases with the total content supply (V):

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

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 formulation captures how total content supply reduces the expected exposure of any *single* post (Iyer and Katona, 2016). It 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, 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). The linear specification keeps creators' utilities affine in P_c, enabling closed-form equilibrium solutions. It offers a tractable way to capture this congestion property that higher aggregate activity lowers per-post reach. Classical congestion models (e.g., Acemoglu and Ozdaglar, 2007; Johari et al., 2010) adopt smooth, monotone mappings with the same logic. To verify that our qualitative results are not driven by this linear approximation, Appendix A.3 re-solves the model numerically using

[5] Allowing moderation on reactions would not materially alter our qualitative results. Moderating negative reactions effectively removes *some* negative engagement that involves uncivil behavior. Incorporating reaction-level moderation would therefore resemble a reduction in $|\omega|$, i.e., the discouraging effect from negative engagement, and the core polarizing mechanism we highlight would continue to hold under the relevant parameter ranges.

three other bounded alternatives, e.g. inverse, exponential, and logistic reach functions, and confirms that all yield qualitatively identical equilibrium patterns.

3. User Engagement $((NE^{in}(i,t) + NE^{out}(i,t))$: If a post is seen, it may generate engagement from both ideological in-group readers ($NE^{in}(i,t)$) and out-group readers ($NE^{out}(i,t)$). These are calculated as expected responses from all reader types:

$$NE^{in}(i,t) = \underbrace{\lambda_i}_{\text{in-group share}} \cdot [\underbrace{P_l(r=i, c=it)}_{\text{prob. of like}} + \omega \underbrace{P_{dl}(r=i, c=it)}_{\text{prob. of dislike}}], \tag{4}$$

$$NE^{out}(i,t) = \underbrace{\lambda_{-i}}_{\text{out-group share}} \cdot [P_l(r=-i, c=it) + \omega P_{dl}(r=-i, c=it)], \tag{5}$$

where P_l and P_{dl} denote, from the perspective of a creator c of type (i,t) , the probabilities that their content receives a like or dislike, respectively, from the in-group ($r = i$) or out-group ($r = -i$) readers, whose respective share is denoted by λ_i and λ_{-i}.[6] Together with the engagement probabilities from Table 2, our specification implies the "hallmarks" of affective polarization, characterized by an emotional divide of *in-group favoritism* and *out-group animosity* (Lerman et al., 2024). That is, creators, even toxic ones, tend to get more positive reactions from ideologically aligned readers (in-group favoritism) and more negative reactions from ideologically opposing readers (out-group animosity). The parameter $\omega \in [-1, 0)$ captures the discouraging effect of negative engagement on content creation. Equivalently, $|\omega|$ represents the creator's degree of 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 4.4.

Finally, 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 each user type that creates content equals the probability of creation implied by their utility. Formally, the following condition must hold for all user types:

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

[6]Here, the λ_i and λ_{-i} can be computed based on Table 1.

In equilibrium, all creators share a common belief about $P_c(i,t)$, which enters the content reach term $R(\cdot)$ and must be internally consistent with the realized outcome. As long as $P_c(\cdot) \in [0,1]$, we have $R(\cdot) \in [0,1]$, ensuring equilibrium existence via Brouwer's fixed-point theorem.

In the next section, we discuss how moderation shapes content creation of each types and their respective welfare.

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 total share of toxic users is not too large ($x \leq \frac{2}{3}$). 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. Proofs for all subsequent result and propositions are also included in the Appendix.

Lemma 1 (**Affective Polarization without Group Imbalance**)**.** *When ideological groups are balanced ($\delta = 0$), the effect of moderation depends entirely on affective polarization (α). For all $i \in \{A, B\}$, $\beta \in [0,1]$, $\tau_A \in (0,1)$:*

(a) Neutral Environment: When $\alpha = 0$, moderation unambiguously increases non-toxic content creation while reduces survived toxic content across both ideological groups:

$$\frac{\partial P_c(i,NT)}{\partial \beta} > 0, \quad \frac{\partial [S(T) \cdot P_c(i,T)]}{\partial \beta} < 0,$$

(b) Polarized Environment: When $\alpha > 0$ and creators are highly averse to negative feedback

($\omega \leq -\frac{1}{2}$), stronger affective polarization can sometimes make moderation counterproductive:

$$\frac{\partial P_c(i, NT)}{\partial \beta} \begin{cases} < 0 & \textit{if } \alpha > \alpha_1(\omega) \equiv \frac{2+\omega}{1-\omega}, \\ \geq 0 & \textit{otherwise}, \end{cases} \qquad \frac{\partial [S(T) \cdot P_c(i, T)]}{\partial \beta} < 0.$$

In the first benchmark case without affective polarization ($\alpha = 0$), readers evaluate content solely by its toxicity rather than ideology. From a creator's perspective, expected engagement therefore depends only on whether their post is toxic or non-toxic, making the effect of moderation symmetric across ideological groups. Removing toxic posts increases the content reach of surviving material because fewer posts compete for reader attention. This broader reach encourages non-toxic creators from both groups to produce more. Meanwhile, the total volume of toxic content that survives moderation declines monotonically.

When affective polarization emerges ($\alpha > 0$), it changes creation incentives. Readers begin to display in-group favoritism and out-group animosity: they react more favorably to ideologically aligned content and more negatively to opposing views. As a result, moderation's impact depends critically on the degree of affective polarization. With moderate α, creation incentives remain roughly balanced across groups. However, when affective polarization is high and creators are highly averse to negative feedback, moderation can become counterproductive: even as it reduces surviving toxic posts, it can discourage the creation of non-toxic users, thereby shrinking the pool of civil content that moderation aims to promote. Conditional on exposure, users' aversion to negative engagement (ω) amplifies the harm of animosity relative to the benefit of favoritism, whether from in-group or out-group members; As α increases, out-group tolerance weakens while their animosity intensifies; once α exceeds a critical threshold $\alpha_1(\omega)$, the expected disutility from out-group animosity outweighs the utility from in-group favoritism, which is amplified by increased content reach to both ideological groups due to moderation. Consistent with this mechanism, Thomas et al. (2022) report that roughly 22% of creators self-censor content (about themselves or their beliefs) to avoid negative reactions, illustrating how anticipated backlash can deter creation.

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. We can show that the total volume of toxic content that survives moderation continues to decline monotonically, i.e., $\frac{\partial [S(T) P_c^*(i,T)]}{\partial \beta} < 0$, consistent with Lemma 1. In the proposition below, we therefore focus on how moderation changes the incentives of *non-toxic* creators, who play a central role in maintaining long-term content quality and user retention on the platform.

Proposition 1 (Spiral of Silence Equilibrium). *For $\beta \in [0, 1]$, $\tau_A \in (0, 1)$, moderating toxic content affects the equilibrium content creation among non-toxic creators from ideological groups A and B differently. Specifically, three distinct regions emerge:*

(a) ***Universal Suppression**: If $\alpha > \bar{\alpha}$, moderation reduces non-toxic content creation from both groups, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} < 0, \frac{\partial P_c^*(B,NT)}{\partial \beta} < 0$.*

(b) ***Universal Empowerment**: If $\alpha < \underline{\alpha}$, moderation increases non-toxic content creation from both groups, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} > 0, \quad \frac{\partial P_c^*(B,NT)}{\partial \beta} > 0$.*

(c) ***Polarized Creation**: If $\underline{\alpha} \leq \alpha \leq \bar{\alpha}$, moderation polarizes creation: majority group creates more whereas minority group creates less, i.e., $\frac{\partial P_c^*(A,NT)}{\partial \beta} \geq 0, \frac{\partial P_c^*(B,NT)}{\partial \beta} \leq 0$.*

Specifically, $\underline{\alpha} = \frac{\omega+2}{(1+2\delta)(1-\omega)}$ and $\bar{\alpha} = \frac{\omega+2}{(1-2\delta)(1-\omega)}$.

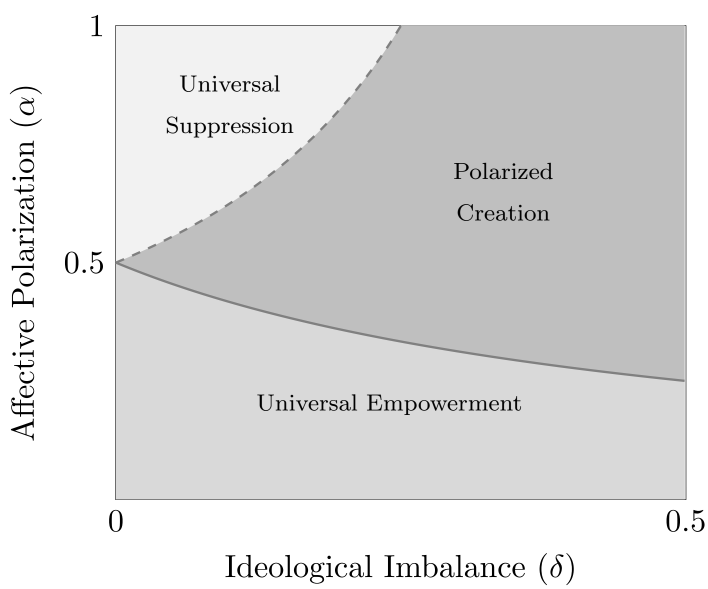


Figure 2: Effect of Moderation on Non-toxic Content Creation ($\omega = -1$).

Note: $\alpha \in [0, 1]$ reflects the degree of affective polarization, i.e., the weight users place on ideology relative to toxicity (higher α means stronger aversion to opposing ideology). $\delta \in [0, 0.5]$ captures ideological imbalance across users, ranging from 0 (equal group sizes) to 0.5 (one group dominates the entire market). The dashed line represents the upper threshold, $\overline{\alpha}$, and the solid line represents the lower threshold, $\underline{\alpha}$.

Proposition 1 shows that when one ideological group dominates the population, moderation no longer affects creators uniformly across ideological lines, for any relative share of toxic users across groups ($\forall \tau_A \in (0,1)$). Instead, its impact depends critically on both the degree of affective polarization (α) and the extent of ideological imbalance (δ). As illustrated in Figure 2, three distinct equilibrium regions emerge. When affective polarization is weak ($\alpha < \underline{\alpha}$), we are in the "universal empowerment" region: moderation enhances the reach of non-toxic content, and the benefits outweigh the costs for non-toxic creators across both groups. This logic echoes that of Lemma 1(a). 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 animosity, ultimately reducing content creation from both sides. This rationale aligns with Lemma 1(b). Between these two extremes lies the "polarized creation" region, which arises under moderate affective polarization ($\underline{\alpha} \leq \alpha \leq \overline{\alpha}$). Here, moderation amplifies existing ideological imbalances: creators from the ideological 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 content reach of non-toxic posts from both groups. However, the resulting engagement dynamics differ sharply: ideological majority creators anticipate more in-group favoritism, while minority creators expect more out-group animosity, which can be strong enough to outweigh the support from their own smaller in-group base.

The final result from Proposition 1 echoes the classic "spiral of silence" effect in social psychology, first described by Noelle-Neumann (1974) as a process through which majority voices dominate public discourse while minority voices fall silent, "increasingly establishing the [majority] opinion as the prevailing one" (p. 44). Traditionally, this spiral is interpreted as a psychological reaction to social pressure. Our model reveals a different, equilibrium-based logic: content moderation can *induce* rather than enforce silence in equilibrium. By changing creator incentives through anticipated content reach, even ideologically neutral moderation could discourage creation from ideological minorities and generate the appearance of bias without any explicit censorship.

Our findings also shed new lights on a long-standing debate: while some commentators attribute higher takedown rates to ideological "censorship," some empirical studies often point instead to "toxicity asymmetry," where certain ideological groups host more toxic content (Haimson et al., 2021). We offer an alternative explanation. Even when the number of toxic users is identical across groups, stricter moderation can *differentially* suppress or amplify non-toxic content, driven entirely by supply-side responses to expected engagement. Thus, the observed asymmetry, where one side appears to "lose" more content, may not reflect inherent toxicity. Rather, it may emerge endogenously from how moderation policies interact with ideological composition and creator incentives.

In this sense, what critics call "censorship" and what empirical studies term "toxicity asymmetry" could be seen as two sides of the same equilibrium process: one that transforms externally neutral moderation into self-censorship.

Building on Proposition 1, 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.

For clarity and simplicity, in the following proposition and extension sections, 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. Our setup allows us to examine how ideological imbalance (δ) and affective polarization (α) shape moderation outcomes, independent of any ideological differences in toxicity.[7] In doing so, we also shed new light on the above debate over whether ideological differences in toxicity actually matter.

Proposition 2 (**Welfare Redistribution**). *Stricter moderation (larger β) enlarges readers' welfare gap across two ideology groups. That is, for any $\alpha > 0, \delta > 0$, we have:*

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

Proposition 2 shows that moderating toxic content redistributes reader welfare across ideological groups. The intuition naturally follows Proposition 1: 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 1 and 2 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

[7] 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.

posts gain greater exposure. Majority-group creators benefit from this shift but do not internalize the negative spillovers, namely, increased out-group animosity 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 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 content reach and engagement can systematically disadvantage ideological minorities. This concern is further complicated 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 group-level fairness, not just accuracy.

In the next section, we develop several important extensions to our baseline model. They serve to further illustrate the limitations of moderation policies that disregard *relational* externalities.

4 Extension

In this section, we explore four extensions of the baseline model. The first introduces content personalization that tailors exposure to ideologically aligned audiences. The second adds a group of ideologically neutral users who are non-partisan and care only about content toxicity. We examine these two extensions first because they are often viewed as quick remedies to the distortions implied by the baseline model: reducing cross-group exposure or neutralizing some users' ideology. Yet, as we show in this section, their effects are more nuanced than these intuitive prescriptions suggest.

By contrast, the last two extensions focus on the alternative motivations for toxic users, first as readers who derive utility from seeing toxic content directed at ideological opponents, and then as creators who value negative engagement. These variations introduce richer behavioral complexity while demonstrating the robustness of our main results. In both cases, the qualitative patterns from the baseline model remain intact, suggesting that our core mechanism does not depend on the specific assumption about toxic users' motivations.

4.1 Content Personalization

Our baseline model assumes that a creator's post is displayed uniformly across the platform such that the expected reach for a single post is identical across two partisan groups. In practice, however, social media platforms frequently personalize exposure based on users' ideological affinity (González-Bailón et al., 2023; Eg et al., 2023). To capture this feature, we extend the model by allowing *differential* content reach and engagement between ideological in-group and out-group audiences. This extension clarifies how the platform's personalization design interacts with moderation outcomes.

We begin from the reader's perspective. A reader of ideology i is exposed to both in-group ($i' = i$) and out-group ($i' = -i$) content of toxicity type $t \in \{NT, T\}$. Now the respective exposure probabilities are given by:

$$P_r^{in}(i,t) = \frac{\lambda(i,t)S(t)P_c(i,t)}{\sum_t \lambda(i,t)S(t)P_c(i,t) + (1-\phi)\lambda(-i,t)S(t)P_c(-i,t)}; \tag{7}$$

$$P_r^{out}(i,t) = \frac{(1-\phi)\lambda(-i,t)S(t)P_c(-i,t)}{\sum_t \lambda(i,t)S(t)P_c(i,t) + (1-\phi)\lambda(-i,t)S(t)P_c(-i,t)}, \tag{8}$$

where $\phi \in [0, 1]$ calibrates how much personalization filters out cross-group exposure. When $\phi = 0$, content reach of both groups is fully symmetric as in the baseline model; higher values of ϕ indicate stronger personalization toward the creator's in-group audience via reduced cross-group exposure.

Accordingly, the creator's expected utility is modified as follows:

$$U^c(i,t) = S(t) \cdot \Big[R^{in} \cdot NE^{in}(i,t) + R^{out} \cdot \, NE^{out}(i,t)\Big],$$

where the reach of a single post to in-group and out-group readers is modified as follows:[8]

$$R^{in} = 1 - V, \quad R^{out} = (1-\phi)(1-V); \quad V = \sum_{(i,t)\in\Theta} \lambda(i,t) \cdot S(t) \cdot P_c(i,t). \tag{9}$$

The modified terms, R^{in} and R^{out}, refer to the content reach to in-group and out-group readers, respectively, where cross-ideological reach is restricted by the degree of personalization ϕ. All other aspects of the model remain unchanged.

In the baseline model, we show that minority creators may withdraw when moderation intensifies, as their content faces stronger out-group animosity. This raises a natural question: could content personalization, by exposing users to less ideologically opposing content, help mitigate this spiral of silence? To examine this possibility, we focus on how personalization changes creation incentives for minority creators. The result is summarized in the following proposition.

Proposition 3. ***(content personalization and moderation)***

(a) When $\phi < \underline{\phi} = \frac{4\delta(1-\omega)+4\omega+2}{3\omega(1-2\delta)}$, our main results regarding non-toxic user's content creation ($\frac{\partial P_c^*(A,NT)}{\partial\beta}$ and $\frac{\partial P_c^*(B,NT)}{\partial\beta}$) remain qualitatively the same.

(b) Under a given moderation policy ($\beta > 0$), personalization may decrease or increase content creation by non-toxic minority creators. In particular, if $\alpha < \underline{\alpha}(\beta,\phi,x,\delta)$, $\frac{\partial P_c^*(B,NT)}{\partial\phi} < 0$, otherwise, $\frac{\partial P_c^*(B,NT)}{\partial\phi} > 0$. Meanwhile, under insufficiently strict moderation ($\beta < \underline{\beta}(\alpha,\delta,\phi,x)$), the surviving minority-toxic content increases with personalization, i.e., $\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\phi} > 0$.

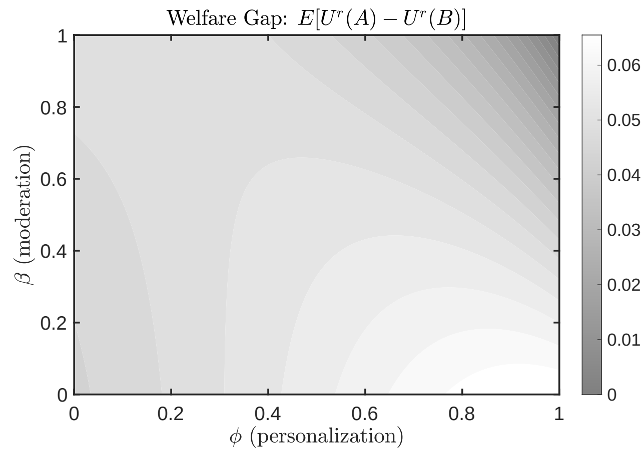
Proposition 3 shows that when personalization remains moderate ($\phi < \underline{\phi}$), our main qualitative results continue to hold. Meanwhile, it also suggests that personalization and moderation are neither perfect substitutes nor complete remedies for the externalities identified in the baseline model. Moderation removes toxic content directly, while personalization redirects exposure toward ideologically aligned readers. For minority creators, these two forces operate in tension: personalization

[8]Note that per-post reach R^{in}, R^{out} and realized exposure shares P_r^{in}, P_r^{out} operate at different aggregation levels and represent distinct perspectives on content exposure, though they remain internally consistent. Specifically, $R(\cdot)$ reflects the ex-ante (expected) reach of a *single* surviving post from the creator's perspective, capturing competition for limited attention, whereas $Pr(\cdot)$ denotes the ex-post share of *total* exposures attributed to content of a given type from the reader's perspective.

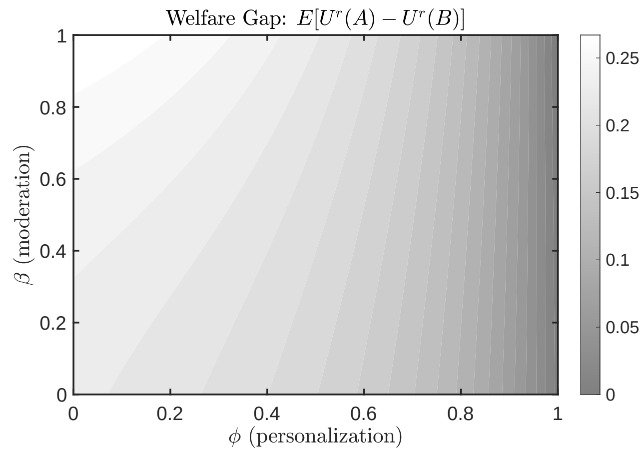
provides only conditional relief for minority non-toxic users, while weakening the disciplining effect of moderation on minority-toxic ones.

Specifically, a high degree of personalization introduces a distinct trade-off for non-toxic creators. On one hand, personalization shields minority creators from the out-group animosity that moderation amplifies in the baseline model. By concentrating exposure within their own ideological group, minority creators anticipate less negative engagement and relatively more positive engagement. This protection can offset the discouragement caused by moderation, leading some non-toxic minority creators to maintain or even increase their creation. On the other hand, personalization also narrows their potential reach: posts are now shown primarily to in-group readers. This reduced audience size weakens the incentive to create, particularly when affective polarization is low and in-group favoritism carries limited emotional payoff. The balance between these two effects determines the slope of minority non-toxic creation with respect to moderation. When affective polarization is mild ($\alpha < \underline{\alpha}$), the loss of reach dominates, so personalization discourages minority non-toxic creation ($\frac{\partial P_c^*(B,NT)}{\partial \phi} < 0$). When affective polarization is high, the effect of protective engagement prevails, and personalization can instead encourage minority creation ($\frac{\partial P_c^*(B,NT)}{\partial \phi} > 0$). Although such personalized exposure may appear beneficial, it still carries the well-documented concern of "echo chambers," where algorithmic curation confines users within ideologically homogeneous environments (Cinelli et al., 2021; Huang et al., 2024; González-Bailón et al., 2023). Beyond this established concern, however, our model identifies an additional class of risk: when the moderation is insufficiently strict ($\beta < \underline{\beta}$), greater personalization also increases the survival of minority-toxic content ($\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial \phi} > 0$). By promoting exposure through sympathetic in-group readers, toxic minority creators are motivated to create more as they anticipate less backlash, enabling more of their posts to persist despite moderation.

To further illustrate how moderation and personalization jointly shape reader welfare, we conduct a numerical analysis under low ($\alpha = 0.2$) and high ($\alpha = 0.9$) affective polarization. The results, summarized in Figure 3, reveal that moderation and personalization are neither perfect substitutes nor perfect complements. As shown in Figure 3a, when both moderation and affective polarization are low, personalization can exacerbate welfare inequality, likely due to reduced cross-group reach. In contrast, Figure 3b presents a different pattern: under high affective polarization, personalization helps narrow the welfare gap created by moderation, primarily by mitigating cross-group animosity when it matters most, though this effect diminishes at high levels of moderation intensity.



(a) $\alpha = 0.2, \delta = 0.1$



(b) $\alpha = 0.9, \delta = 0.1$

Figure 3: $\mathbb{E}[U^r(A) - U^r(B)]$ over degree of personalization (ϕ) and moderation intensity (β)

Taken together, our analysis shows that moderation and personalization interact in non-linear ways. When affective polarization is strong, moderate personalization can complement strict moderation by protecting minority creators from out-group animosity. But when moderation is weak, the same personalization revives toxic creation among the minority. In less polarized environments, personalization not only offers in-group comfort to toxic creators but weakening cross-group validation for non-toxic ones. In other words, even as content personalization protects the minority creators from out-group animosity, it can simultaneously induce a higher level of toxic content and sometimes even exacerbate the "spiral of silence" discussed earlier. This creates a subtler form of polarization beyond the standard echo-chamber explanations. These patterns highlight the importance of evaluating policy fairness at different levels of affective polarization (α) and jointly over both moderation and personalization, rather than treating each dimension in isolation.

From a practical standpoint, depending on their user base and mission, platforms face distinct optimal combinations of (β, ϕ). Highly polarized platforms (e.g., X/Twitter, Truth Social) may benefit from moderate personalization that limits cross-group hostility while maintaining strict moderation to minimize in-group toxicity. More heterogeneous or knowledge-oriented communities (e.g., Reddit, Wikipedia) should resist strong personalization to preserve content diversity and reduce risks of echo-chamber.

4.2 Ideological Neutral Users

In the baseline model, all users are partisan ($i \in \{A, B\}$). In practice, some users on the platform may not have strong ideology preference. According to the survey by Pew Research Center (2024), around 35% of the registered voters say they are independent. This extension incorporates that possibility by introducing an ideologically neutral group of users ($i = N$). Their utility function as a reader r seeing a post c is simply given by

$$U^r(r = N, c = i't') = V(t) + \varepsilon_r.$$

Here, we assume that the neutral readers only care about the toxicity aspect of content. In terms of content creation, for simplicity, we rule out the possibility that neutral users post toxic content. Accordingly, the user type space is defined as $\Theta = \{(N, NT)\} \cup (\{A, B\} \times \{T, NT\})$. Let $\lambda_N \in (0, 1)$ denote the mass of neutral users. Accordingly, the mass of A-group and B-group are given by $(1-\lambda_N)(\frac{1}{2}+\delta)$ and $(1-\lambda_N)(\frac{1}{2}-\delta)$, respectively. Given the addition of neutral (non-toxic) content, the utility function of partisan readers $i \in \{A, B\}$ is now modified as follows:

$$U^r(r = i, c = i't') = \alpha \cdot H(i, i') + (1 - \alpha) \cdot V(t) + \varepsilon_r,$$

where $H(i, i')$ is given by:

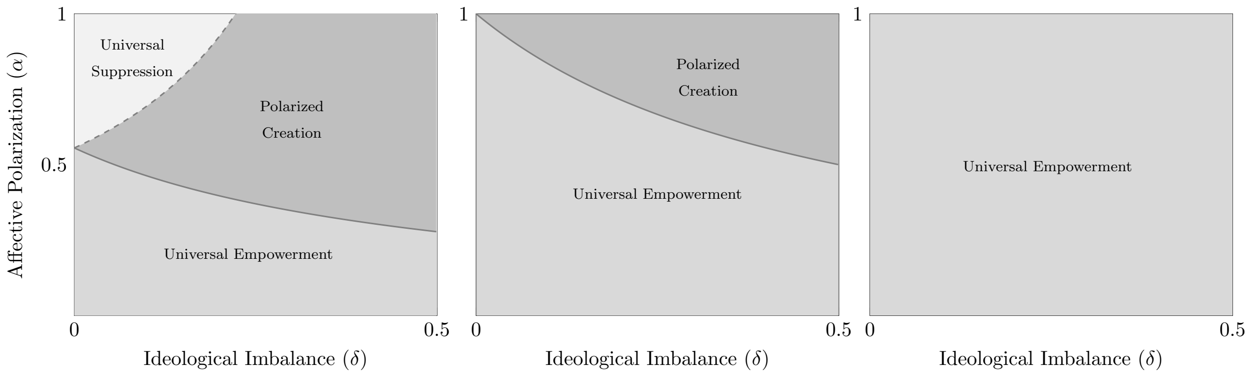
$$H(i, i') = \begin{cases} 0, & \text{if } i = i', \\ -\frac{1}{2}, & \text{if } i' = N, \\ -1, & \text{otherwise.} \end{cases}$$

The horizontal dimension $H(i, i')$ takes three values depending on which group the content is from. We assume that reader prefer content from the same ideology group to neutral group to the opposing group. This specification retains the feature of in-group favoritism and out-group animosity from the baseline model, while adding that partisan readers exhibit (symmetrically) moderate tolerance toward ideologically neutral content. All other aspects of the model remain unchanged. The result is summarized in the following proposition.

Proposition 4. ***(Ideological Neutral Users)***

(a) Moderation increases neutral content if and only if the share of neutral users is large enough, i.e., if $\lambda_N < \lambda_0 \equiv 2 - \frac{3}{1-\omega}$, then $\frac{\partial P_c^*(N,NT)}{\partial \beta} < 0$; otherwise, $\frac{\partial P_c^*(N,NT)}{\partial \beta} \geq 0$.

(b) Moderation induces the following three equilibrium regions at varying shares of neutral users:



(a) $\lambda_N = 0.1$ (b) $\lambda_N = 0.5$ (c) $\lambda_N = 0.9$

Figure 4: Effect of Moderation in the Presence of Neutral Users ($\omega = -1$; $\lambda_N \in \{0.1, 0.5, 0.9\}$)

- *Universal Suppression: when $\lambda_N < \underline{\lambda_N}$, moderation suppresses the creation of both partisan groups, i.e.,* $\frac{\partial P_c^*(A,NT)}{\partial \beta} < 0, \quad \frac{\partial P_c^*(B,NT)}{\partial \beta} < 0.$
- *Polarized Creation: when $\underline{\lambda_N} \leq \lambda_N \leq \overline{\lambda_N}$, moderation polarizes the creation of partisan groups, i.e.,* $\frac{\partial P_c^*(A,NT)}{\partial \beta} \geq 0, \quad \frac{\partial P_c^*(B,NT)}{\partial \beta} \leq 0.$
- *Universal Empowerment: when $\lambda_N > \overline{\lambda_N}$, moderation empowers the creation of both partisan groups, i.e.,* $\frac{\partial P_c^*(A,NT)}{\partial \beta} > 0, \quad \frac{\partial P_c^*(B,NT)}{\partial \beta} > 0.$

Specifically, $\overline{\lambda_N} \equiv 1 - \frac{\omega+2}{\alpha(1-\omega)(1+2\delta)}$ *and* $\underline{\lambda_N} \equiv 1 - \frac{\omega+2}{\alpha(1-\omega)(1-2\delta)}$.

Proposition 4 shows how the presence of ideologically neutral users affects the impact of moderation on partisan users' content creation incentives. Neutral readers systematically reward non-toxic content, diluting out-group animosity and generating no ideology-based backlash. Proposition 4(a) shows that when such users make up only a small share of the platform, they may become even less active as moderation intensifies. The logic parallels the spiral of silence: with a limited audience base, neutral creators effectively become the new minority, where appreciation from their small in-group is outweighed by partisan indifference or mild hostility. As a result, they are more likely to remain silent despite a declining level of toxicity on the platform. Only once their share exceeds a critical threshold do neutral creators become more active under stricter moderation. In that regime, neutral users not only soften the tension between the two partisan groups but also crowd out their reach. Consequently, the incentives of non-toxic partisan creators depend on the trade-off between expected reach and engagement.

Proposition 4(b) further shows that when the neutral-user share is very small ($\lambda < \underline{\lambda_N}$), outcomes resemble those in the baseline model. As their share rises to moderate levels ($\underline{\lambda_N} \leq \lambda \leq \overline{\lambda_N}$), their role as an "animosity absorber" outweighs the crowd-out effect on majority creators, eliminating the region of universal suppression but still generating a spiral-of-silence equilibrium under high affective polarization. Finally, when neutral users dominate the platform ($\lambda > \underline{\lambda_N}$), cross-group animosity among partisans largely dissipates, and the outcome converges to that in Lemma 1(a). Figure 4 illustrates how the two partisan groups respond to moderation under different shares of neutral users.

From a policy perspective, our findings in this extension suggest that encouraging ideological neutrality among only a few users can deepen their isolation, as they lose audience without changing the broader climate of hostility. Only when neutrality becomes common enough to reshape the overall climate does moderation begin to promote content creation rather than silence.

4.3 Targeted Toxicity Homophily

In our baseline model, toxicity is treated strictly as a vertical trait, i.e., it is uniformly disliked by everyone. Whereas in practice, some toxic creators may derive satisfaction from toxicity directed at their ideological opponents. Empirical studies find that ideologically aligned users react more favorably to toxic attacks *on* the out-group (e.g., Yu et al., 2024; Lerman et al., 2024). To accommodate this interpretation while maintaining the distinction between ideology and toxicity, we extend the reader–creator utility to allow what we refer to as *targeted toxicity homophily*: toxic readers value toxicity when it is targeted toward the opposing group, i.e., when it is aligned with their own ideology. Formally, the reader $r = (i, t)$'s utility from a creator $c = (i', t')$'s post is now given by:

$$U^r(r = it, c = i't') = \alpha \cdot H(i, i') + (1 - \alpha) \cdot \mathcal{T}(it, i't') + \varepsilon_r,$$

where

$$H(i, i') = \begin{cases} 0, & i = i', \\ -1, & i \neq i', \end{cases} \qquad \mathcal{T}(it, i't') = \begin{cases} 0 & t' = NT \\ -1, & t' = T, t = NT \\ \dfrac{\kappa\alpha}{1-\alpha} \cdot \mathbf{1}[i = i'] - 1, & t' = T, t = T \end{cases},$$

and $\kappa \in (0, 1)$ measures the strength of this targeted homophily. When $\kappa = 0$, the model reverts to the baseline in which toxicity uniformly reduces all reader utility; for $\kappa > 0$, toxic readers obtain additional utility from toxic content produced by same-ideology creators, i.e., $\mathcal{T}(iT, iT) = \frac{\kappa\alpha}{1-\alpha} - 1$. In contrast, when they encounter toxic content from the opposing group, their utility remain the same as in the baseline model, i.e. $\mathcal{T}(iT, -iT) = -1$. In other words, toxic readers continue to dislike toxicity directed toward their own side but gain satisfaction when toxicity aligns with their in-group ideology. For non-toxic readers, they continue to dislike toxic content regardless of the ideology associated with the creator, consistent with the baseline model. This extension captures the identity-affirming function of hostility frequently observed in online discourse, where "punching the other side" garners approval within the in-group. In this sense, $\mathcal{T}(it, i't')$ reflects the relational nature of toxicity itself within polarized environment.

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

Proposition 5. *Compared to the baseline model, when toxic users value toxic contents from the same group,*

- *the main results regarding non-toxic user's content creation ($\frac{\partial P_c^*(A,NT)}{\partial \beta}$ and $\frac{\partial P_c^*(B,NT)}{\partial \beta}$) and welfare inequality ($\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta}$) remain qualitatively unchanged, $\forall \beta \in [0,1]$.*
- *the (surviving) toxic content increases with the degree of targeted toxicity homophily. That is, $\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial \kappa} > 0$, $\forall i \in \{A, B\}$, $\kappa \in (0,1)$.*

Proposition 5 shows that under targeted toxicity homophily, our main qualitative results remain unchanged: the three equilibrium regions from the baseline model persist. This is because targeted homophily directly shapes how toxic readers engage with toxic content, thereby influencing the incentives of toxic creators. In contrast, it affects *non-toxic* creators only indirectly through its impact on the equilibrium feedback loop. Despite this indirect channel, the expected reach of non-toxic posts still increases with moderation. Thus, the core intuition of the baseline model continues to hold.

The key difference from the baseline model lies in the behavior of toxic rather than non-toxic creators. The homophily term raises toxic creators' expected positive engagement (and dampens expected negative engagement) toward same-side toxic content. Consequently, toxic creators are further incentivized to produce toxic posts.

4.4 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 be motivated by negative responses. For instance, Buckels et al. (2014) find that certain toxic individuals derive pleasure from eliciting anger or outrage.

To capture this heterogeneity in content creation incentives, 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 $\omega \in [-1, 0)$. Recall that the penalty (or reward) from negative engagement, captured by ω, enters the in-group and out-group engagement terms as follows:

$$NE^{in}(i,t) = \lambda_i \cdot [P_l(r = i, c = it) + \omega(t) \cdot P_{dl}(r = i, c = it)],$$
$$NE^{out}(i,t) = \lambda_{-i} \cdot [P_l(r = -i, c = it) + \omega(t) \cdot P_{dl}(r = -i, c = it)] \,.$$

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

Proposition 6. *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* $\frac{\partial P_c(B,NT)}{\partial \beta}$) *and welfare inequality* ($\frac{\partial \mathbb{E}[(U^r(A)-U^r(B)]}{\partial \beta}$) *remain qualitatively unchanged* $\forall \beta \in [0,1]$. *In addition, the equilibrium level of (surviving) toxic content is higher than the main model.*

This extension confirms the robustness of our main findings. The intuition is as follows. When toxic users value negative engagement, their incentives for content creation are directly strengthened. Their impact on non-toxic creators' content creation is only through the change in content supply and corresponding changes in expected reach. We show that moderation consistently raises their expected reach and, conditional on reach, their expected engagement remains the same. Consequently, the strategic environment faced by non-toxic users remains largely unchanged. As a result, key outcomes, such as their content creation and welfare inequality, continue to move in the same direction as in the baseline model. As for toxic users, by contrast, valuing negative engagement strengthens their motivation to produce toxic content, leading to a higher level of toxic content in equilibrium.

5 Conclusion

This paper offers a strategic perspective on moderating toxicity in online platforms, showing that content moderation is not only about what gets removed, but also shapes what gets created. Importantly, we show that even when moderation targets toxicity alone, their effects may not be ideologically neutral: stricter enforcement can encourage ideological majority creation at the expense of silencing minority. These asymmetries arise not from explicit bias but from structural externalities in content governance. As a result, the same policy can silence, stimulate, or polarize creation depending on the structure of affective divide within its user base.

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 across the board. 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.

To highlight the core mechanism, our model deliberately abstracts from two additional forces: endogenous user exit and advertising incentives. Under certain conditions, incorporating them would likely reinforce rather than overturn our results. Allowing readers to leave the platform when their utility falls below a threshold would disproportionately drive minority attrition, shrinking their audience base and amplifying ideological imbalance, thereby accelerating the spiral toward "polarized creation." Likewise, we omit advertisers to focus on consumer behavioral responses to moderation, but the insights extend naturally to a profit-maximizing platform: when moderation is guided by engagement metrics, it can still magnify inequality in creation and welfare across ideological groups, even under rules designed to be neutral. Future research can formally test these conjectures in settings where both user retention and revenue incentives are endogenous.

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 stricter moderation. We hope future research will examine these predictions in field or experimental settings to guide evidence-based content governance.

References

Acemoglu, D. and Ozdaglar, A. (2007). Competition and efficiency in congested markets. *Mathematics of operations research*, 32(1):1–31.

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.

Berman, R. and Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316.

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

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*.

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

Cima, L., Trujillo, A., Avvenuti, M., and Cresci, S. (2024b). The great ban: Efficacy and unintended consequences of a massive deplatforming operation on reddit. In *Companion Publication of the 16th ACM Web Science Conference*, pages 85–93.

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.

Craker, N. and March, E. (2016). The dark side of facebook®: The dark tetrad, negative social potency, and trolling behaviours. *Personality and Individual Differences*, 102:79–84.

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.

Eg, R., Tønnesen, Ö. D., and Tennfjord, M. K. (2023). A scoping review of personalized user experiences on social media: The interplay between algorithms and human factors. *Computers in Human Behavior Reports*, 9:100253.

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*.

González-Bailón, S., Lazer, D., Barberá, P., Zhang, M., Allcott, H., Brown, T., Crespo-Tenorio, A., Freelon, D., Gentzkow, M., Guess, A. M., et al. (2023). Asymmetric ideological segregation in exposure to political news on facebook. *Science*, 381(6656):392–398.

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.

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.

Iyer, G. and Katona, Z. (2016). Competing for attention in social communication markets. *Management Science*, 62(8):2304–2320.

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*.

Johari, R., Weintraub, G. Y., and Van Roy, B. (2010). Investment and market structure in industries with congestion. *Operations Research*, 58(5):1303–1317.

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.

Lerman, K., Feldman, D., He, Z., and Rao, A. (2024). Affective polarization and dynamics of information spread in online networks. *npj Complexity*, 1(1):8.

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

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. (2023a). The effects of online content moderation: Evidence from president trump's account deletion. *Available at SSRN 4296306*.

Müller, K. and Schwarz, C. (2023b). 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.

Pew Research Center (2021). The state of online harassment. *Pew Research Center*, January 13, 2021. Accessed: June 30, 2025.

Pew Research Center (2024). The partisanship and ideology of american voters. Accessed October 29, 2025.

Rabbani, M. G. and Pusch, N. (2025). Explaining the victim-offender overlap of cyberbullying using low self-control and parental bonds. *Crime & Delinquency*, 71(10):3219–3243.

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.

Rieder, B. and Skop, Y. (2021). The fabrics of machine moderation: Studying the technical, normative, and organizational structure of perspective api. *Big Data & Society*, 8(2):20539517211046181.

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.

Thomas, K., Kelley, P. G., Consolvo, S., Samermit, P., and Bursztein, E. (2022). "it's common and a part of being a content creator": Understanding how creators experience and cope with hate and harassment online. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, pages 1–15.

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.

Yu, X., Wojcieszak, M., and Casas, A. (2024). Partisanship on social media: In-party love among american politicians, greater engagement with out-party hate among ordinary users. *Political Behavior*, 46(2):799–824.

A Appendix

A.1 Equilibrium Characterization

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

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

where

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

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

$$
R^* = -\frac{8(\beta x+1)}{M} \tag{A-2}
$$

We compute the welfare inequality using the following expression:

$$
\mathbb{E}\left[U^r(A)-U^r(B)\right] = \sum_{(i',t')\in\{A,B\}\times\{NT,T\}} Pr(i',t')\cdot\left[U^r\left(A,i't'\right)-U^r(B,i't')\right] \tag{A-3}
$$

where $Pr(i,t) = \frac{\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}{\sum_{i\in\{A,B\},t\in\{T,NT\}}\lambda(i,t)\cdot S(t)\cdot P_c^*(i,t)}$ is the probability of a reader viewing content of type (i,t), given the equilibrium level of content creation.

Based on Table 2, we also derive the exact expressions for in-group and out-group net engagement (NE) for each type of creators in the baseline model, shown below:

$$
\begin{aligned}
NE^{in}(A, NT) &= \frac{1}{8}(2\delta + 1)(\omega + 2) \\
NE^{in}(B, NT) &= \frac{1}{8}(1 - 2\delta)(\omega + 2) \\
NE^{in}(A, T) &= \frac{1}{8}(2\delta + 1)(2\alpha(1 - \omega) + 3\omega) \\
NE^{in}(B, T) &= \frac{1}{8}(2\delta - 1)(2\alpha(\omega - 1) - 3\omega) \\
NE^{out}(A, NT) &= \frac{1}{8}(1 - 2\delta)(2\alpha(\omega - 1) + \omega + 2) \\
NE^{out}(B, NT) &= \frac{1}{8}(2\delta + 1)(2\alpha(\omega - 1) + \omega + 2) \\
NE^{out}(A, T) &= \frac{3}{8}(1 - 2\delta)\omega \\
NE^{out}(B, T) &= \frac{3}{8}(2\delta + 1)\omega.
\end{aligned}
$$

A.2 Proofs

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

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

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

From equation (A-2), we take derivative with respect to β and obtain:

$$\frac{\partial R^*}{\partial \beta} = \frac{4xN_1(x)}{\left(-\alpha\left(4\delta^2-1\right)(\omega-1)+x(-\alpha(\omega-1)((\beta-2)\beta(4\delta\tau_A-2\delta+1)+2)+(3(\beta-2)\beta+2)\omega-2)+\omega+10\right)^2},$$

where

$$\begin{aligned}N_1(x) &= \alpha(\omega-1)(\beta(8\delta\tau_A-4\delta+2)-4\delta(\delta+2\tau_A-1)-1)-6\beta\omega \\ &\quad + x\left(\alpha(\omega-1)\left(\beta^2(4\delta\tau_A-2\delta+1)-2\right)-3\beta^2\omega+2\omega-2\right)+7\omega+10.\end{aligned}$$

The denominator is a square and positive; Since $4x > 0$, as long as $N_1(x)$ defined above is positive, we have $\frac{\partial R^*}{\partial \beta} > 0$.

We can easily see that $N_1(x)$ is a linear function of x. Hence,

$$\min_{x\in[0,\frac{2}{3}]} N_1(x) = \min\left\{N_1(0), N_1(\frac{2}{3})\right\}.$$

It is sufficient to show $\min_{x\in[0,\frac{2}{3}]} N_1(x) > 0$ to prove $\frac{\partial R^*}{\partial \beta} > 0$.

Step 1.1: $N(0)$

$$\begin{aligned}N_1(0) &= \alpha(\omega-1)(\beta(8\delta\tau_A-4\delta+2)-4\delta(\delta+2\tau_A-1)-1)+(7-6\beta)\omega+10 \\ &\geq \alpha(1-\omega)(-8\delta\tau_A+4\delta-2+4\delta(\delta+2\tau_A-1)+1)-7+10 \\ &= \alpha(1-\omega)(4\delta^2-1)-7+10 \\ &\geq -2\alpha-7+10 \\ &> 0\end{aligned}$$

Here, the first inequality follows from $\beta(8\delta\tau_A-4\delta+2)-4\delta(\delta+2\tau_A-1)-1$ increases with β, as $8\delta\tau_A-4\delta+2 > 0$. The second last inequality follows from $\delta \in [0, \frac{1}{2}]$ and $\omega \in [-1, 0)$ and the last inequality follows $\alpha \in [0, 1]$. As a result, we have $N_1(0) > 0$.

Step 1.2: $N_1(\frac{2}{3})$

$$
\begin{aligned}
N_1(\frac{2}{3}) &= \frac{1}{3}\left(\alpha(\omega-1)\left(\beta^2(8\delta\tau_A-4\delta+2)+6\beta(4\delta\tau_A-2\delta+1)-12\delta(\delta+2\tau_A-1)-7\right)\right)\\
&\quad+\frac{1}{3}\left((25-6\beta(\beta+3))\omega+26\right).
\end{aligned}
$$

We can easily see that $N_1(\frac{2}{3})$ is a linear function of ω. Hence,

$$
\min_{\omega\in[-1,0]} N_1(\frac{2}{3}) = \min\left\{N_1(\frac{2}{3})|_{\omega=0}, N_1(\frac{2}{3})|_{\omega=-1}\right\}.
$$

It is sufficient to show $\min_{\omega\in[-1,0]} N_1(\frac{2}{3}) > 0$ to prove $N_1(\frac{2}{3}) > 0$.

$$
\begin{aligned}
N_1(\frac{2}{3})|_{\omega=0} &= \frac{1}{3}\left(\alpha\left(\beta^2(-8\delta\tau_A+4\delta-2)-6\beta(4\delta\tau_A-2\delta+1)+12\delta(\delta+2\tau_A-1)+7\right)+26\right)\\
&= \frac{\alpha}{3}((-2\beta^2-6\beta)(4\delta\tau_A-2\delta+1)+12\delta(\delta+2\tau_A-1)+7)+\frac{26}{3}\\
&\geq \frac{1}{3}(-16+4+26)\\
&> 0.
\end{aligned}
$$

The first inequality follows from $\delta(4\tau_A-2)+1>0$ and $12\delta(\delta+2\tau_A-1)+7\geq 4$ is minimized when $\delta=\frac{1}{2}(1-2\tau_A)$ and $\tau_A=0$, as well as $\delta\in[0,\frac{1}{2}]$, and $\alpha\in[0,1]$.

$$
\begin{aligned}
N_1(\frac{2}{3})|_{\omega=-1} &= \frac{1}{3}\left(2\alpha\left(\beta^2(-8\delta\tau\mathrm{A}+4\delta-2)-6\beta(4\delta\tau\mathrm{A}-2\delta+1)+12\delta(\delta+2\tau\mathrm{A}-1)+7\right)+6\beta(\beta+3)+1\right)\\
&= \frac{1}{3}(2\alpha((-2\beta^2-6\beta)(4\delta\tau_A-2\delta+1)+12\delta(\delta+2\tau_A-1)+7)+6\beta(\beta+3)+1)\\
&\geq \frac{1}{3}(2(-2(2\beta^2+6\beta)+4)+6\beta^2+18\beta+1)\\
&\geq \frac{1}{3}(9-2\beta(3+\beta))\\
&> 0
\end{aligned}
$$

Similar to the above analysis, the first inequality follows from $12\delta(\delta+2\tau_A-1)+7\geq 4$ and $0\leq(4\delta\tau_A-2\delta+1)\leq 2$. The second inequality follows $\beta\in[0,1]$. As a result, we have $N_1(\frac{2}{3})\geq \min_{\omega\in[-1,0]} N_1(\frac{2}{3}) = \min\left\{N_1(\frac{2}{3})|_{\omega=0}, N_1(\frac{2}{3})|_{\omega=-1}\right\} > 0$.

Since both $N_1(0)>0$ and $N_1(\frac{2}{3})>0$, we have $N_1(x)>0$ for all $x\in(0,\frac{2}{3}]$. Therefore, the numerator $4xN_1(x)>0$, and the denominator is positive. Hence, we have $\frac{\partial R^*}{\partial\beta}>0$.

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

$$
\begin{aligned}
&\frac{d[(1-\beta)R^*]}{d\beta} = -R^*+(1-\beta)\frac{\partial R^*}{\partial\beta}\\
&= \frac{4N_2(x,\omega)}{\left(-\alpha\left(4\delta^2-1\right)(\omega-1)+x(-\alpha(\omega-1)((\beta-2)\beta(4\delta\tau_A-2\delta+1)+2)+(3(\beta-2)\beta+2)\omega-2)+\omega+10\right)^2}
\end{aligned}
$$

where

$$
\begin{aligned}
N_2(x,\omega) = &\alpha\left(4\delta^2-1\right)(\omega-1) - x^2(\alpha(\omega-1)(\beta(\beta\delta(4\tau_A-2)+\beta-4)+2)+\beta((4-3\beta)\omega-4)-2\omega+2) \\
&- x\left(\alpha(\omega-1)\left(\beta^2(4\delta\tau_A-2\delta+1)-4\beta\delta(2\delta+2\tau_A-1)+4\delta(\delta+2\tau_A-1)-1\right)\right) \\
&- x\left(\beta((8-3\beta)\omega+20)-5\omega-12\right)-\omega-10
\end{aligned}
$$

Here, the denominator is a positive square. Hence, it is sufficient to show $\max_{\omega\in[-1,0]} N_2(x,\omega) < 0$ to prove $\frac{d[(1-\beta)R^*]}{d\beta} < 0$.

We can easily see that $N_2(x,\omega)$ is a linear function of ω. Hence,

$$
\max_{\omega\in[-1,0]} N_2(x,\omega) = \max\{N_2(x,-1), N_2(x,0)\}.
$$

Step 2.1: $N_2(x,0)$

$$
\begin{aligned}
N_2(x,0) = &-4\alpha\delta^2+\alpha+x^2(\alpha(\beta(\beta\delta(4\tau_A-2)+\beta-4)+2)+4\beta-2) \\
&+ x\left(\alpha\left(\beta^2(4\delta\tau_A-2\delta+1)-4\beta\delta(2\delta+2\tau_A-1)+4\delta(\delta+2\tau_A-1)-1\right)-20\beta+12\right)-10.
\end{aligned}
$$

Here, $N(x,0)$ is decreasing in $\beta\in[0,1]$. That is,

$$
\begin{aligned}
\frac{\partial N_2(x,0)}{\partial\beta} &= 2x(4\alpha(\beta-1)\delta\tau_A+\alpha(1-2\delta)(\beta+2\delta)) \\
&\quad + 2x(x(\alpha(\beta\delta(4\tau_A-2)+\beta-2)+2)-10) \\
&\le 2x(-10+1+2x) \\
&< 0
\end{aligned}
$$

The first inequality follows from $\alpha(1-2\delta)(\beta+2\delta)\le 1$, $4\alpha(\beta-1)\delta\tau_A\le 0$, and $\beta\delta(4\tau_A-2)+\beta-2\le 0$. The second inequality follows from $x\in(0,\frac{2}{3}]$. Therefore, $N_2(x,0)$ decreases with β.

Thus, we have

$$
\begin{aligned}
N_2(x,0) &\le N_2(x,0)|_{\beta=0} \\
&= -4\alpha\delta^2+\alpha+2(\alpha-1)x^2+x(\alpha(4\delta(\delta+2\tau_A-1)-1)+12)-10 \\
&= 4\alpha(x-1)\delta^2+4\alpha x\,(2\tau_A-1)\,\delta+\alpha+x(-\alpha+2(\alpha-1)x+12)-10 \\
&< \frac{2}{9}(10\alpha-13) \\
&< 0.
\end{aligned}
$$

The first inequality follows from $4\delta(\delta+2\tau_A-1)-1\le 2$, for $\delta\in[0,\frac{1}{2}]$ and $\tau_A\in[0,1]$. Moreover, $N_2(x,0)|_{\beta=0}$ is an increasing function of τ_A and a quadratic function of δ with the coefficient of δ^2

being $(4\alpha(x-1))$. Thus, the maximum value is achieved at $\delta = \frac{1}{2}$ and $x = \frac{2}{3}$. The last inequality follows from $\alpha \in [0,1]$.

Step 2.2: $N_2(x,-1)$

$$N_2(x,-1) = \alpha\left(2-8\delta^2\right) + x^2(2\alpha(\beta(\beta\delta(4\tau_A-2)+\beta-4)+2)+\beta(8-3\beta)-4)$$
$$+ x\left(2\alpha\left(\beta^2(4\delta\tau_A-2\delta+1)-4\beta\delta(2\delta+2\tau_A-1)+4\delta(\delta+2\tau_A-1)-1\right)-3\beta(\beta+4)+7\right)-9.$$

Hence, $N_2(x,-1)$ is decreasing in β for $\beta \in [0,1]$. That is,

$$\begin{aligned}\frac{\partial N_2(x,-1)}{\partial\beta} &= 2x(8\alpha(\beta-1)\delta\tau_A + 2\alpha(1-2\delta)(\beta+2\delta)) \\ &+ 2x(-3(\beta+2)+x(2\alpha(\beta(4\delta\tau_A-2\delta+1)-2)-3\beta+4)) \\ &\leq 2x(2-6+4x) \\ &< 0.\end{aligned}$$

The second last inequality follows from $\beta(4\delta\tau_A-2\delta+1)-2 \leq 0$, $\alpha(1-2\delta)(\beta+2\delta) \leq 1$, $8\alpha(\beta-1)\delta\tau_A \leq 0$ and $\beta \in [0,1]$. The last inequality follows $x \in (0, \frac{2}{3}]$. So $N_2(x,-1)$ decreases with β. Therefore,

$$\begin{aligned}N_2(x,-1) \leq& N_2(x,-1)|_{\beta=0} \\ &= \alpha\left(2-8\delta^2\right) + 4(\alpha-1)x^2 \\ &+ x(8\alpha\delta(\delta+2\tau_A-1)-2\alpha+7)-9 \\ &= 8(x-1)\alpha\delta^2 + 8x\alpha\delta + 2\alpha + x(-2\alpha+4(\alpha-1)x+7)-9 \\ &\leq \frac{1}{9}(46\alpha-55) \\ &< 0.\end{aligned}$$

The second last inequality follows from $4(\alpha-1)x^2 \leq 0$, and $\alpha \in [0,1]$. Moreover, $N_2(x,-1)|_{\beta=0}$ is a quadratic function of δ with the coefficient of δ^2 being negative, and an increasing function of τ_A. Therefore, the maximum is achieved at $\hat{\delta} = \frac{x}{2-2x}$. The last inequality follows $\alpha \in [0,1]$. Hence, we have $N(x,-1) < 0$. Since both $N_2(x,0) < 0$ and $N_2(x,-1) < 0$, we have $N_2(x) < 0$ for all $x \in (0, \frac{2}{3}]$. Therefore, the numerator $4xN_2(x) < 0$, and the denominator is positive. Hence, we have $\frac{\partial[(1-\beta)R^*]}{\partial\beta} < 0$.

□

Proof of Lemma 1(a) 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 two claims outlined in Lemma 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:

$$\frac{\partial P_c^*(i,NT)}{\partial \beta} = \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \left(NE^{in}(i,NT) + NE^{out}(i,NT)\right).$$

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

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

First, according to the equilibrium condition (6), we can have:

$$P_c^*(i,T) = \frac{1}{2} + \frac{1}{2} \cdot S(T) \cdot R^* \cdot \left(NE^{in}(i,T) + NE^{out}(i,T)\right).$$

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

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

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

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

$$\begin{aligned}
\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial \beta} &= -P_c^*(i,T) + (1-\beta) \cdot \frac{\partial P_c^*(i,T)}{\partial \beta} \\
&= -\frac{1}{2} - \frac{1}{2}R^*(1-\beta)\frac{3\omega}{4} + \frac{(1-\beta)}{2}\frac{\partial[(1-\beta)R^*]}{\partial \beta} \cdot \frac{3\omega}{4} \\
&= -\frac{1}{2} - \frac{1}{2}(1-\beta)\frac{3\omega}{4}\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial \beta}\right) \\
&\leq -\frac{1}{2} + \frac{3}{8}(1-\beta)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial \beta}\right).
\end{aligned}$$

Given the expression of R^* and $\frac{\partial[(1-\beta)R^*]}{\partial \beta}$, we can express $(1-\beta)(R^* - \frac{\partial[(1-\beta)R^*]}{\partial \beta})$ as follows: $(1-\beta)(R^* - \frac{\partial[(1-\beta)R^*]}{\partial \beta}) = \frac{4(\beta-1)(\omega(x(-3\beta+(2-3(\beta-2)(\beta-1)\beta)x+3)-2)+2(x-5)((3\beta-1)x+2))}{(x((3(\beta-2)\beta+2)\omega-2)+\omega+10)^2} = \frac{N_1(\omega)}{D_1^2}$.

Here, the denominator is a square of $D_1 = x((3(\beta-2)\beta+2)\omega-2)+\omega+10$, which is affine in x.

Hence,

$$\min_{x \in [0,\frac{2}{3}]} D_1 = \min\{D_1|_{x=0}, D_1|_{x=\frac{2}{3}}\}.$$

We can show that

$$D_1|_{x=0} = 10 + \omega \geq 9$$

and

$$D_1|_{x=\frac{2}{3}} = \frac{1}{3}((6(\beta-2)\beta+7)\omega+26) \geq \frac{25}{3},$$

The inequality follows from $D_1|_{x=\frac{2}{3}}$ increases with ω and β . Therefore, we have the denominator

$$(D_1)^2 \geq (\min\{9, \frac{25}{3}\})^2 \geq \frac{625}{9}.$$

For the numerator, we can write it as $N_1(x)$ where

$$N_1(\omega) = 4(\beta-1)(\omega(x(-3\beta+(2-3(\beta-2)(\beta-1)\beta)x+3)-2)+2(x-5)((3\beta-1)x+2))$$

We can easily see that it is affine in ω. Hence, we have,

$$\max_{\omega\in[-1,0)} N_1(\omega) = \max\{N_1(0), N_1(-1)\}$$

we can show

$$N_1(0) = 8(\beta-1)(x-5)((3\beta-1)x+2)$$

$N_1(0)$ is a quadratic function of β with coefficient of β^2 as negative and the maximum value occur when $\beta = \frac{2}{3} - \frac{1}{3x}$. Hence,

$$N_1(0) \leq \begin{cases} -\frac{8(x-5)(x+1)^2}{3x} \leq 54 & x \in [\frac{1}{2}, \frac{2}{3}] \\ 8(x-5)(x-2) \leq 80 & x \in (0, \frac{1}{2}] \end{cases}$$

And,

$$N_1(-1) = 4(\beta-1)(x(-27\beta+(3\beta((\beta-3)\beta+4)-4)x+11)-18).$$

$N_1(-1)$ is a quadratic function of x with coefficient of x^2 as positive if $x \in [0, \bar{x}]$ and otherwise, the coefficient of x^2 is negative. Hence, we have two cases:

Case 1: if $x \in (0, \bar{x}]$,

$N_1(-1)$ is a quadratic function with the maximum value occuring at the endpoints of x. That is, $N_1(-1) \leq \max\{N_1(-1)|_{x=0}, N_1(-1)|_{x=\frac{2}{3}}\}$. Here, we can show,

$$N_1(-1)|_{x=0} = 72(1-\beta) \leq 72,$$
$$N_1(-1)|_{x=\frac{2}{3}} = \frac{8}{9}\left(3(2(\beta-4)\beta-13)\beta^2+\beta+56\right) \leq \frac{448}{9}.$$

Case 2: if $x \in (\bar{x}, \frac{2}{3})$

We have $\max N_1(-1) = \max\{N_1(-1)|_{x=0}, N_1(-1)|_{x=\frac{2}{3}}, N_1(-1)|_{x=x^*}\}$, where $x^* = \frac{27\beta-11}{6\beta((\beta-3)\beta+4)-8}$. We have shown the expressions of $N_1(-1)|_{x=0}$, and $N_1(-1)|_{x=\frac{2}{3}}$ above,

$$N_1(-1)|_{x=x^*} = -\frac{(\beta-1)(27\beta(\beta(8\beta+3)+10)-167)}{3\beta((\beta-3)\beta+4)-4} \leq \frac{448}{9}.$$

Therefore, given $N_1(\omega) \leq \max_{\omega\in[-1,0)} N(\omega) = 80$ and $D^2 \geq \frac{625}{9}$, we have

$$\begin{aligned}
\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial\beta} &\leq -\frac{1}{2} + \frac{3}{8}(1-\beta)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial\beta}\right) \\
&\leq -\frac{1}{2} + \frac{3}{8}\frac{N_1(\omega)}{D_1^2} \\
&< 0.
\end{aligned}$$

□

Proof of Lemma 1 (b) 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

$$\begin{aligned}
\frac{dP_c^*(i,NT)}{d\beta} =& \frac{1}{2}\cdot\frac{dR^*}{d\beta}\cdot\left(NE^{in}(i,NT) + NE^{out}(i,NT)\right) \\
=& \frac{1}{2}\cdot\frac{dR^*}{d\beta}\cdot\left(\frac{\omega+\alpha(\omega-1)+2}{4}\right).
\end{aligned}$$

According to Lemma A.1, we know that $\frac{dR^*}{d\beta} > 0$. 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

$$\begin{aligned}
\frac{dP_c^*(i,T)}{d\beta} =& \frac{1}{2}\frac{d[(1-\beta)R^*]}{d\beta}\cdot\left(NE^{in}(i,T) + NE^{out}(i,T)\right) \\
=& \frac{1}{2}\frac{d[(1-\beta)R^*]}{d\beta}\cdot\left(\frac{\alpha(1-\omega)+3\omega}{4}\right).
\end{aligned}$$

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

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

$$\begin{aligned}
\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial\beta} &= -P_c^*(i,T) + (1-\beta)\cdot\frac{\partial P_c^*(i,T)}{\partial\beta} \\
&= -\frac{1}{2} - \frac{R^*}{2}\cdot(1-\beta)\left(\frac{\alpha(1-\omega)+3\omega}{4}\right) + \frac{(1-\beta)}{2}\frac{\partial[(1-\beta)R^*]}{\partial\beta}\cdot\left(\frac{\alpha(1-\omega)+3\omega}{4}\right) \\
&= -\frac{1}{2} - \frac{1}{2}(1-\beta)\left(\frac{\alpha(1-\omega)+3\omega}{4}\right)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial\beta}\right).
\end{aligned}$$

Here, we have two cases:

Case 1: if $\left(\frac{\alpha(1-\omega)+3\omega}{4}\right) > 0$,

Together with $R^* > 0$ and $\frac{\partial[(1-\beta)R^*]}{\partial\beta} < 0$, we will have $\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial\beta} < 0$ always holds.

Case 2: If $\left(\frac{\alpha(1-\omega)+3\omega}{4}\right) < 0$,

We have $\left(\frac{\alpha(1-\omega)+3\omega}{4}\right) \geq -\frac{3}{4}$ and it increases with α and ω.

Given the expression of R^* and $\frac{\partial[(1-\beta)R^*]}{\partial\beta}$, the effect of α and ω on them is secondary through the changes in $P_c(i,t)$ and $NE^{in}(i,t)$ and $NE^{out}(i,t)$. That means the effects will be multiplied by $\frac{1}{2}\cdot\lambda(i,t) < \frac{1}{2} < 1$. Hence, $\left(\frac{\alpha(1-\omega)+3\omega}{4}\right)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial\beta}\right)$ should increase with α and ω and the minimum value should occur when $\alpha = 0$ and $\omega = -1$.

Hence we have

$$
\begin{aligned}
&(1-\beta)\left(\frac{\alpha(1-\omega)+3\omega}{4}\right)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial\beta}\right)\\
\leq&\left[(1-\beta)\left(\frac{\alpha(1-\omega)+3\omega}{4}\right)\left(R^* - \frac{\partial[(1-\beta)R^*]}{\partial\beta}\right)\right]_{\omega=-1,\alpha=0}\\
=&\frac{N_2(\alpha,\omega)}{D_2^2}.
\end{aligned}
$$

Following the same logic as in Lemma 1(a), we can show that $\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial\beta} < 0$. □

Proof of Proposition 1 According to the equilibrium condition (6), we take derivative of $P_c^*(A,NT)$ w.r.t. β and obtain:

$$
\begin{aligned}
\frac{\partial P_c^*(A,NT)}{\partial\beta} =& \frac{1}{2}\cdot\frac{\partial R^*}{\partial\beta}\cdot\left(NE^{in}(A,NT) + NE^{out}(A,NT)\right)\\
=&\frac{1}{2}\cdot\frac{\partial R^*}{\partial\beta}\cdot\left[\frac{\omega+\alpha(\omega-1)(1-2\delta)+2}{4}\right],
\end{aligned}
$$

which is negative if and only if $\alpha > \bar{\alpha} \equiv \frac{\omega+2}{(1-2\delta)(1-\omega)}$ as Lemma A.1 shows that $\frac{dR^*}{d\beta} > 0$.

Taking derivative of $P_c^*(B,NT)$ wrt β, we obtain:

$$
\begin{aligned}
\frac{\partial P_c^*(B,NT)}{\partial\beta} =& \frac{1}{2}\cdot\frac{\partial R^*}{\partial\beta}\cdot\left(NE^{in}(B,NT) + NE^{out}(B,NT)\right)\\
=&\frac{1}{2}\cdot\frac{\partial R^*}{\partial\beta}\cdot\left[\frac{\omega+\alpha(\omega-1)(2\delta+1)+2}{4}\right],
\end{aligned}
$$

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

It is easy to verify that $\bar{\alpha} > \underline{\alpha}$ always holds for $\omega \in [-1,0)$ and $\delta \in [0,\frac{1}{2}]$. 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 1. □

Proof of Proposition 2 As indicated on page 16, we henceforth focus on the case $\tau_A = \tau_B = \frac{1}{2}$. Based on the definition of welfare inequality in equation (A-3), we can rewrite it as follows:

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

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,

$$\frac{\partial \mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta} = \alpha \left(\frac{\partial Pr(A, NT)}{\partial \beta} - \frac{\partial Pr(B, NT)}{\partial \beta} + \frac{\partial Pr(A, T)}{\partial \beta} - \frac{\partial Pr(B, T)}{\partial \beta} \right). \tag{A-4}$$

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

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

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-4), we have

$$\frac{\partial \mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta} = \alpha \cdot \left(\frac{\partial [\frac{\Delta P_c(NT) + \Delta P_c(T)}{1 - R^*}]}{\partial \beta} \right).$$

Let

$$G(\alpha) \equiv \frac{\partial \mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta}, \quad K(\alpha) \equiv \frac{\partial [\frac{\Delta P_c(NT) + \Delta P_c(T)}{1 - R^*}]}{\partial \beta}.$$

It is easy to verify that

$$G(\alpha) = \alpha \cdot K(\alpha), \quad K(0) = 0.$$

Differentiating w.r.t. α,

$$\frac{\partial G}{\partial \alpha} = K(\alpha) + \alpha \cdot \frac{\partial K(\alpha)}{\partial \alpha},$$

We can show $\frac{\partial G}{\partial \alpha} \geq 0$, so $G(\alpha)$ is increasing in α. Since $G(\alpha) = 0$, we conclude $G(\alpha) \geq 0$ for all $\alpha \in [0, 1]$ and hence $\frac{\mathbb{E}[U^r(A) - U^r(B)]}{\partial \beta} \geq 0$. $\square$

Proof of Proposition 3 As indicated on page 16, we henceforth restrict attention to the case $\tau_A = \tau_B = \frac{1}{2}$. Under this assumption, the earlier bound $x \in (0, \frac{2}{3}]$, derived without specifying τ_A and τ_B, can be relaxed to $x \in (0, \frac{3}{4}]$ (a less restrictive condition), which we adopt in the subsequent proofs.

Part (a):

The proof logic is similar to Lemma A.1 and Proposition 1.

Step 1. Show $\frac{dR^*_{in}}{d\beta} > 0$ and $\frac{dR^*_{out}}{d\beta} > 0$

Recall that

$$\begin{aligned} R_{in} &= 1 - \sum_{t\in\{T,NT\},i\in\{A,B\}} \lambda(i,t)\cdot S(T)\cdot P_c(i,t), \\ R_{out} &= (1-\phi)(1 - \sum_{t\in\{T,NT\},i\in\{A,B\}} \lambda(i,t)\cdot S(T)\cdot P_c(i,t)). \end{aligned}$$

Since $\phi \in [0,1]$, it is sufficient to show $\frac{dR^*_{in}}{d\beta} > 0$.

$$\frac{\partial R^*_{in}}{\partial \beta} = \frac{8x\,N(x)}{\begin{bmatrix} -2\alpha(4\delta^2-1)(\omega-1)(\phi-1) - 4\delta^2\omega\phi - 8\delta^2\phi \\ +2\alpha x(\omega-1)\big((\beta-2)\beta-\phi+2\big) \\ +x(\phi-2)\big((3(\beta-2)\beta+2)\omega-2\big) \\ -2\omega+\omega\phi+2\phi-20 \end{bmatrix}^2}.$$

where

$$\begin{aligned} N(x) = {} & 2\alpha(\omega-1)\left(2\beta+4\delta^2(\phi-1)-\phi-1\right) - 12\beta\omega + 6\beta\omega\phi + 4\delta^2\omega\phi \\ & + 8\delta^2\phi + 2\alpha x(\omega-1)\left(\beta^2+\phi-2\right) \\ & + x(\phi-2)\left(\left(3\beta^2-2\right)\omega+2\right) + 14\omega - 7\omega\phi - 2\phi + 20. \end{aligned}$$

Here, the denominator is a square, and $8x > 0$. So as long as the affine-in-x factor $N(x)$ is positive, we will have $\frac{\partial R^*_{in}}{\partial \beta} > 0$. Hence, it is sufficient to show

$$\min_{x\in[0,\frac{3}{4}]} N(x) = \min\{N(0), N(\frac{3}{4})\} > 0.$$

Step 1.1: $N(0)$

$$\begin{aligned} N(0) = {} & 2\alpha(\omega-1)\left(2\beta+4\delta^2(\phi-1)-\phi-1\right) \\ & + \phi\left(6\beta\omega+4\delta^2(\omega+2)-7\omega-2\right) + 2(7-6\beta)\omega + 20. \end{aligned}$$

Here, $N(0)$ is affine in ϕ, hence $\min N(0) = \min\{N(0)|_{\phi=1}, N(0)|_{\phi=0}\}$.

$$
\begin{aligned}
N(0)|_{\phi=0} &= 2\alpha(\omega-1)\left(2\beta-4\delta^2-1\right)+2(7-6\beta)\omega+20\\
&\geq 2\alpha(\omega-1)\left(2\beta-4\delta^2-1\right)+2(7-6\beta)\omega+20\\
&\geq -2\alpha+14\omega+20\\
&> 0.
\end{aligned}
$$

The third inequality follows $|\omega-1|\leq 2$ and $|2\beta-4\delta^2-1|\leq 1$.

$$
\begin{aligned}
N(0)|_{\phi=1} &= 4\alpha(\beta-1)(\omega-1)-6\beta\omega+4\delta^2(\omega+2)+7\omega+18\\
&> -7+18\\
&> 0
\end{aligned}
$$

The second last inequality in both expressions follows from $\delta\in[0,\frac{1}{2}]$, $\beta\in[0,1]$ and $\alpha\in[0,1]$ and the last inequality in both expressions follows $\omega\in[-1,0)$. Combine these two together, we obtain $N(0)\geq\min\{N(0)|_{\phi=0},N(0)|_{\phi=1}\}>0$.

Step 1.2: $N(\frac{3}{4})$

$$
\begin{aligned}
N(\frac{3}{4}) = &\frac{1}{4}\left(2\alpha(\omega-1)\left(\beta(3\beta+8)+16\delta^2(\phi-1)-\phi-10\right)\right)\\
&+\frac{1}{4}(3\beta(3\beta+8)\omega(\phi-2)+16\delta^2(\omega+2)\phi+68(\omega+1)-34\omega\phi-2\phi).
\end{aligned}
$$

We can show that $N(\frac{3}{4})$ is a linear function in ω. Hence, $\min N(0)=\min\{N(\frac{3}{4})|_{\omega=-1},N(\frac{3}{4})|_{\omega=0}\}$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=0} &= \frac{1}{2}\left(\alpha\left(-\beta(3\beta+8)-16\delta^2(\phi-1)+\phi+10\right)+\left(16\delta^2-1\right)\phi+34\right)\\
&\geq \frac{1}{2}(-\beta(3\beta+8)+10+16\delta^2+34)\\
&\geq \frac{1}{2}(-11+10-1+34)\\
&> 0.
\end{aligned}
$$

The second and third inequalities follow $\beta(3\beta+8)\leq 11$, $\phi(16\delta^2-1)>-1$ and $16\delta^2(\phi-1)>-4$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=-1} &= \alpha\left(-\beta(3\beta+8)-16\delta^2(\phi-1)+\phi+10\right)\\
&\quad -\frac{3}{4}\beta(3\beta+8)(\phi-2)+4\delta^2\phi+8\phi\\
&\geq (-\beta(3\beta+8)+10)+\frac{3}{2}\beta(3\beta+8)\\
&= \frac{1}{2}\beta(3\beta+8)+10\\
&> 0.
\end{aligned}
$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\phi \in [0,1]$ and $\alpha \in [0,1]$, and the last inequality in both expressions follows $\beta \in [0,1]$. Combine these two together, we obtain $N(\frac{3}{4}) \geq \min\{N(\frac{3}{4})|_{\omega=0}, N(\frac{3}{4})|_{\omega=-1}\} > 0$. Since both $N(0) > 0$ and $N(\frac{3}{4}) > 0$, we have $N(x) > 0$ for all $x \in (0, \frac{3}{4})$. Therefore, the numerator $4xN(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial R^*_{in}}{\partial \beta} > 0$.

Step 2: $\frac{\partial P^*_c(i,NT)}{\partial \beta}$

Taking derivative of $P^*_c(A, NT)$ w.r.t. β, we obtain

$$\begin{aligned}\frac{\partial P^*_c(A,NT)}{\partial \beta} =& \frac{1}{2}[\frac{\partial R^*_{in}}{\partial \beta} \cdot NE^{in}(A,NT) + \frac{\partial R^*_{out}}{\partial \beta} NE^{out}(A,NT)] \\ =& \frac{1}{2}\frac{\partial R^*_{in}}{\partial \beta} \cdot [NE^{in}(A,NT) + (1-\phi)NE^{out}(A,NT)] \\ =& \frac{1}{2}\frac{\partial R^*_{in}}{\partial \beta} \cdot \frac{(2\delta-1)(\phi-1)(2\alpha(\omega-1)+\omega+2)+(2\delta+1)(\omega+2)}{8}.\end{aligned}$$

We showed that $\frac{\partial R^*_{in}}{\partial \beta} > 0$, so the equation above is negative if $\alpha > \overline{\alpha_p} \equiv \frac{-2\delta\omega\phi-4\delta\phi-2\omega+\omega\phi+2\phi-4}{2(2\delta-1)(\omega-1)(\phi-1)}$.

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

$$\begin{aligned}\frac{\partial P^*_c(B,NT)}{\partial \beta} =& \frac{1}{2}[\frac{\partial R^*_{in}}{\partial \beta} \cdot NE^{in}(B,NT) + \frac{\partial R^*_{out}}{\partial \beta} NE^{out}(B,NT)] \\ =& \frac{1}{2}\frac{\partial R^*_{in}}{\partial \beta} \cdot [NE^{in}(B,NT) + (1-\phi)NE^{out}(B,NT)] \\ =& \frac{1}{2}\frac{\partial R^*_{in}}{\partial \beta} \cdot \left[-\frac{2\alpha(2\delta+1)(\omega-1)(\phi-1)+(\omega+2)(2\delta\phi+\phi-2)}{8}\right].\end{aligned}$$

We showed that $\frac{\partial R^*_{in}}{\partial \beta} > 0$, so the equation above is negative if $\alpha > \underline{\alpha_p} \equiv -\frac{(\omega+2)(2\delta\phi+\phi-2)}{2(2\delta+1)(\omega-1)(\phi-1)}$. Thus, it is easy to verify that $\overline{\alpha_p} > \underline{\alpha_p}$ 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 3. By comparing the two cutoffs $\overline{\alpha_p}$ and $\underline{\alpha_p}$, we can easily verify that the universal suppression exist when $\overline{\alpha_p} < 1$ which is valid when $\phi < \underline{\phi} = \frac{4\delta(1-\omega)+4\omega+2}{3\omega(1-2\delta)}$.

Part (b):

Step 1: Derive the sign conditions for $\frac{\partial P^*_c(B,NT)}{\partial \phi}$.

$$\frac{\partial P_c^*(B,NT)}{\partial \phi} = \frac{2(\beta x+1)\,N(x)}{\begin{bmatrix} -2\alpha(4\delta^2-1)(\omega-1)(\phi-1)-4\delta^2\omega\phi-8\delta^2\phi \\ +2\alpha x(\omega-1)\big((\beta-2)\beta-\phi+2\big) \\ +x(\phi-2)\big((3(\beta-2)\beta+2)\omega-2\big) \\ -2\omega+\omega\phi+2\phi-20 \end{bmatrix}^2},$$

where

$$\begin{aligned} N(\alpha) &= \alpha^2(\beta-1)^2(2\delta+1)x(\omega-1)^2-\alpha(\omega-1)(2(2\delta+1)(\delta(\omega+2)+4) \\ &\quad +x((\beta-2)\beta(2\delta+1)(\omega-1)-6\delta+\omega-1))-(\omega+2)(\delta(2\delta(\omega+2) \\ &\quad +x((3(\beta-2)\beta+2)\omega-2)+\omega+10)+4) \\ &= C_2\alpha^2+C_1\alpha-C_0. \end{aligned}$$

Here, the denominator is a square, and $2(\beta x+1)>0$. The sign of $\frac{\partial P_c^*(B,NT)}{\partial \phi}$ only depends on the sign of $N(\alpha)$. Note that $N(\alpha)$ is a quadratic function of α. Let coefficient of α^2 be defined as $C_2=(\beta-1)^2(2\delta+1)x(\omega-1)^2>0$.

Let C_1 denote the coefficient of α, given by:

$$\begin{aligned} C_1 &= -(\omega-1)(2(2\delta+1)(\delta(\omega+2)+4)+x((\beta-2)\beta(2\delta+1)(\omega-1)-6\delta+\omega-1)) \\ &> 4(1-\omega)+x(2-\beta)\beta(1-\omega)-x(1-\omega) \\ &> (4+x(2-\beta)\beta-x)(1-\omega) \\ &> 0. \end{aligned}$$

Let C_0 denote the constant, given by:

$$\begin{aligned} C_0 &= (\omega+2)(\delta(2\delta(\omega+2)+x((3(\beta-2)\beta+2)\omega-2)+\omega+10)+4) \\ &> \delta x(3\beta(\beta-2)+2-2)+9+4 \\ &> -\frac{3}{8}(-3+2-2)+13 \\ &> 0. \end{aligned}$$

As a result,

$$\frac{\partial N(\alpha)}{\partial \alpha} = 2C_2\alpha+C_1>0.$$

Therefore, there exists a single threshold $\underline{\alpha}$ such that

$$\begin{cases} \alpha < \underline{\alpha} \;\Rightarrow\; N(\alpha) < 0 \;\Rightarrow\; \dfrac{\partial P_c(B,NT)}{\partial \phi} < 0, \\ \alpha > \underline{\alpha} \;\Rightarrow\; N(\alpha) > 0 \;\Rightarrow\; \dfrac{\partial P_c(B,NT)}{\partial \phi} > 0. \end{cases}$$

Step 2: Derive the sign conditions for $\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\phi}$

$$\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\phi} = \frac{8(1-\beta)^2(\beta x+1)\,N(x)}{\begin{bmatrix} 4\alpha(4\delta^2-1)(\omega-1)(\phi-1)+8\delta^2\omega\phi+16\delta^2\phi+4\omega-2\omega\phi-4\phi+40 \\ -2x\Big(2\alpha(\omega-1)\big((\beta-2)\beta-\phi+2\big)+(\phi-2)\big((3(\beta-2)\beta+2)\omega-2\big)\Big) \end{bmatrix}^2},$$

where

$$\begin{aligned} N(x) =& -\alpha^2(2\delta-1)(\omega-1)^2\left(4\delta^2+x-1\right) \\ &+\alpha(\omega-1)\left((2\delta+1)(1-2\delta)^2(\omega-1)+x\omega(6(\beta-2)\beta\delta+8\delta-1)-2\delta x+x\right) \\ &-3\omega(\delta(2\delta(\omega+2)+x((3(\beta-2)\beta+2)\omega-2)+\omega+10)+4). \end{aligned}$$

Here, the denominator is a square, and $8(1-\beta)^2(\beta x+1) > 0$. The sign of $\frac{\partial P_c^*(B,NT)}{\partial\phi}$ only depends on the sign of $N(x)$, which is affine in x.

$$\min_{x\in[0,\frac{3}{4}]} N(x) \in \{N(0), N(\frac{3}{4})\}.$$

Step 2.1 $N(0)$

$$\begin{aligned} N(0) &= (2\delta+1)\left(-\alpha^2(1-2\delta)^2(\omega-1)^2+\alpha(1-2\delta)^2(\omega-1)^2-3\omega(\delta(\omega+2)+4)\right) \\ &\geq (2\delta+1)(1-2\delta)^2(\omega-1)^2(-\alpha^2+\alpha) \\ &> 0 \end{aligned}$$

Step 2.2 $N(\frac{3}{4})$

$$\begin{aligned} N(\frac{3}{4}) &= -\alpha^2(2\delta-1)(4\delta^2-\frac{1}{4})(\omega-1)^2-3\omega\left(\frac{1}{4}\delta((9(\beta-2)\beta+10)\omega+8\delta(\omega+2)+34)+4\right) \\ &+\frac{1}{4}\alpha(\omega-1)(2\delta((9(\beta-2)\beta+8)\omega+8\delta(2\delta-1)(\omega-1)+1)+\omega-1) \\ &= D_2\beta^2+D_1\beta+D_0 \end{aligned}$$

Here,

$$D_2 = \frac{9}{4}\delta\omega(2\alpha(\omega-1)-3\omega)$$
$$D_1 = \frac{9}{2}\delta\omega(3\omega-2\alpha(\omega-1)) = -2C_2$$
$$\begin{aligned}D_0 &= -\alpha^2(2\delta-1)(\omega-1)^2\left(4\delta^2+x-1\right)+\alpha(\omega-1)\left((2\delta+1)(1-2\delta)^2(\omega-1)+(8\delta-1)x\omega-2\delta x+x\right)\\ &\quad -3\omega(\delta(2\delta(\omega+2)+2x(\omega-1)+\omega+10)+4)\end{aligned}$$

Since $N(\frac{3}{4})$ is a quadratic function of β, we can rewrite $N(\frac{3}{4})$ as follows:

$$N(\frac{3}{4}) = D_2(1-\beta)^2 + D_0 - D_2.$$

The minimum of $N(\frac{3}{4})$ depends on the sign of D_2. More importantly, the extreme value occurs when $\beta = 1$. That means, it is either increasing or decreasing in β for all $\beta \in [0,1]$ and $\min N(\frac{3}{4})$ occurs when $\beta = 1$ or $\beta = 0$. In particular,

$$\begin{aligned}N(\frac{3}{4})|_{\beta=0} &= D_0|_{x=\frac{3}{4}}\\ N(\frac{3}{4})|_{\beta=1} &= (-D_2+D_0)|_{x=\frac{3}{4}}\end{aligned}$$

Therefore,

$$\min N(\frac{3}{4}) = \begin{cases} D_0 & \text{if } D_2 \le 0 \\ D_0 - D_2 & \text{if } D_2 > 0. \end{cases}$$

Moreover, $D_2 < 0$ implies $D_2 \le 0 \Rightarrow 2\alpha(\omega-1)-3\omega \le 0 \Rightarrow \omega \le \frac{2\alpha}{2\alpha-3}$. In this case, we can check the sign of D_0.

$$\begin{aligned}D_0 &= \frac{1}{4}\left(-\alpha^2(2\delta-1)\left(16\delta^2-1\right)(\omega-1)^2+\alpha(\omega-1)(2\delta(8\delta(2\delta-1)(\omega-1)+8\omega+1)+\omega-1)\right)\\ &\quad+\frac{1}{2}\left(-3\omega(\delta(4\delta(\omega+2)+5\omega+17)+8)\right)\\ &> \frac{1}{4}\left(-\alpha^2(2\delta-1)\left(16\delta^2-1\right)(\omega-1)^2+\alpha(\omega-1)(2\delta(8\delta(2\delta-1)(\omega-1)+8\omega+1)+\omega-1)\right)\\ &\quad+\frac{1}{2}\left(2\alpha(\delta(4\delta(\omega+2)+5\omega+17)+8)\right)\\ &> \frac{1}{4}(-4\alpha^2-2\alpha(-1)+8\alpha)\\ &= -\alpha^2+\frac{5}{2}\alpha\\ &\ge 0.\end{aligned}$$

The first inequality follows $\omega \le \frac{2\alpha}{2\alpha-3}$ so that $-3\omega > 2\alpha(1-\omega) > 2\alpha$ and $(\delta(4\delta(\omega+2)+5\omega+17)+8) > 0$. The second last inequality follows $\delta \in [0, \frac{1}{2}]$ and $\omega \in [-1, 0)$, and the last inequality follows from $\alpha \in [0,1]$.

Therefore, we have the follow cases:

(1) if $D_2 \le 0$ and $D_0 > 0$, then $N(\frac{3}{4}) > \min_{\beta\in[0,1]} N(\frac{3}{4}) = N(\frac{3}{4}|_{\beta=0}) > 0$ for all $\beta \in [0,1]$.

(2) if $D_2 > 0$ and $D_0 \ge D_2$, then $N(\frac{3}{4}) > \min_{\beta\in[0,1]} N(\frac{3}{4}) = N(\frac{3}{4}|_{\beta=1}) > 0$ for all $\beta \in [0,1]$.

(3) if $D_2 > 0$ and $D_0 < D_2$, there exists a unique $\beta = \underline{\beta} = 1 - \sqrt{1 - D_0/D_2}$ such that

$$\begin{cases} \beta < \underline{\beta} \;\Rightarrow\; N(x) > 0 \;\Rightarrow\; \frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\phi} > 0, \\ \\ \beta > \underline{\beta} \;\Rightarrow\; N(x) < 0 \;\Rightarrow\; \frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\phi} < 0. \end{cases}$$

$\square$

Proof of Proposition 4 We prove the two claims outlined in Proposition 4 below.

The proof involves two steps:

Step 1: show $\frac{\partial R^*}{\partial \beta} > 0$ in this extended model.

$$\frac{\partial R^*}{\partial \beta} = \frac{4(1-\lambda_N)x\,N(x)}{\begin{bmatrix} -(\lambda_N-1)^2(\omega-1)\Big(\alpha\big(4\delta^2+((\beta-2)\beta+2)x-1\big)+1\Big) \\ +(1-\lambda_N)\big(x\big((3(\beta-2)\beta+2)\omega-2\big)+\omega-1\big)+\omega+10 \end{bmatrix}^2}.$$

where

$$\begin{aligned} N(x) =& (7-6\beta)\omega + (\lambda_N-1)^2(\omega-1)\left(-4\alpha\delta^2+\alpha+\alpha\left(\beta^2-2\right)x-1\right) \\ &+(1-\lambda_N)\left((\omega-1)(2\alpha(\beta-1)+1)+x\left(\left(2-3\beta^2\right)\omega-2\right)\right)+10 \end{aligned}$$

Here, the denominator is a square, and $4x(1-\lambda_N) > 0$. So as long as the affine-in-x factor $N(x)$ is positive, we will have $\frac{\partial R^*}{\partial \beta}$. Since $N(x)$ is a linear function of x, it is sufficient to show

$$\min_{x\in[0,\frac{3}{4}]} N(x) = \min\{N(0), N(\frac{3}{4})\} > 0.$$

Step 1.1: $N(0)$

$$\begin{aligned} N(0) = &(1-\lambda_N)(\omega-1)(2\alpha(\beta-1)+1)+(7-6\beta)\omega+10 \\ &+(\lambda_N-1)^2(\omega-1)\left(-4\alpha\delta^2+\alpha-1\right). \end{aligned}$$

Here, $N(0)$ is affine in ω, hence $N(0) \ge \min\{N(0)|_{\omega=-1}, N(0)|_{\omega=0}\}$.

$$
\begin{aligned}
N(0)|_{\omega=0} &= \alpha(\lambda_N - 1)\left(2\beta + 4\delta^2(\lambda_N - 1) - \lambda - 1\right) + \lambda_N(\lambda_N - 1) + 10 \\
&> 0, \\
N(0)|_{\omega=-1} &= 2\alpha(\lambda_N - 1)\left(2\beta + 4\delta^2(\lambda_N - 1) - \lambda_N - 1\right) \\
&+ 6\beta + 2(\lambda_N - 1)\lambda_N + 3 \\
&\geq 2\alpha(1 - \lambda_N)(\lambda_N + 1) + 2\lambda_N(\lambda_N - 1) + 3 \\
&\geq -2\lambda_N(1 - \lambda_N) + 3 \\
&> 0
\end{aligned}
$$

Here, $N(0)|_{\omega=-1}$ increases with β. So the minimum value of $N(0)|_{\omega=-1}$ occurs when $\beta = 0$. The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\beta \in [0,1]$ and $\alpha \in [0,1]$ and the last inequality in both expressions follows $\lambda_N \in (0,1)$. Combine these two together, we obtain $N(0) \geq \min\{N(0)|_{\omega=0}, N(0)|_{\omega=-1}\} > 0$.

Step 1.2: $N(\frac{3}{4})$

$$
\begin{aligned}
N(\frac{3}{4}) = &\frac{1}{4}(\lambda_N - 1)^2(\omega - 1)\left(\alpha\left(3\beta^2 - 16\delta^2 - 2\right) - 4\right) \\
&- \frac{1}{4}(\lambda_N - 1)\left(8\alpha(\beta - 1)(\omega - 1) + \left(10 - 9\beta^2\right)\omega - 10\right) + (7 - 6\beta)\omega + 10.
\end{aligned}
$$

We can show that $N(\frac{3}{4})$ is a linear function in ω. Hence, $\min N(0) \geq \min\{N(\frac{3}{4})|_{\omega=-1}, N(\frac{3}{4})|_{\omega=0}\}$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=0} &= -\frac{1}{4}(\lambda_N - 1)^2\left(\alpha\left(3\beta^2 - 16\delta^2 - 2\right) - 4\right) \\
&+ \frac{1}{2}(\lambda_N - 1)(4\alpha(\beta - 1) + 5) + 10 \\
&\geq -\frac{1}{4}(\lambda_N - 1)^2(-3) + \frac{5}{2}(\lambda_N - 1) + 10 \\
&> 0. \\
N(\frac{3}{4})|_{\omega=-1} &= -\frac{1}{2}(\lambda_N - 1)^2\left(\alpha\left(3\beta^2 - 16\delta^2 - 2\right) - 4\right) \\
&+ \frac{1}{4}(\lambda_N - 1)\left(16\alpha(\beta - 1) - 9\beta^2 + 20\right) + 6\beta + 3 \\
&\geq -\frac{1}{2}(1 - \lambda_N)^2(-3) + \frac{1}{4}(\lambda_N - 1)(-9 + 20) + 3 \\
&= \frac{3}{2}(1 - \lambda_N)^2 + \frac{11}{4}(\lambda_N - 1) + 3 \\
&> 0.
\end{aligned}
$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\beta \in [0,1]$ and $\alpha \in [0,1]$. The last inequality in both expressions follows $\lambda_N \in (0,1)$. Combine these two together, we obtain

$N(\frac{3}{4}) \geq \min\{N(\frac{3}{4})|_{\omega=0}, N(\frac{3}{4})|_{\omega=-1}\} > 0$.

Since both $N(0) > 0$ and $N(\frac{3}{4}) > 0$, we have $N(x) > 0$ for all $x \in (0, \frac{3}{4}]$. Therefore, the numerator $4x(1-\lambda_N)N(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial R^*}{\partial \beta} > 0$.

Step 2: Derive the sign of $\frac{\partial P_c^*(i,NT)}{\partial \beta}$ for $i \in \{A, B, N\}$.

Taking derivative of $P_c^*(N, NT)$ w.r.t. β, we obtain

$$\begin{aligned}\frac{\partial P_c^*(N,NT)}{\partial \beta} =& \frac{1}{2}\frac{\partial R^*}{\partial \beta} \cdot \big(NE^{in}(N,NT) + NE^{out}(N,NT)\big)\\ =& \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \left[\frac{\lambda_N(1-\omega)+2\omega+1}{4}\right]\end{aligned}$$

We have shown that $\frac{dR^*}{d\beta} > 0$. Hence, $\frac{\partial P_c^*(N,NT)}{\partial \beta}$ is negative if $\lambda_N < \lambda_0 \equiv 2 - \frac{3}{1-\omega}$.

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

$$\begin{aligned}\frac{\partial P_c^*(A,NT)}{\partial \beta} =& \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \big(NE^{in}(A,NT) + NE^{out}(A,NT)\big)\\ =& \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \left[\frac{\alpha(2\delta-1)(\lambda_N-1)(\omega-1)+\omega+2}{4}\right]\end{aligned}$$

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

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

$$\begin{aligned}\frac{\partial P_c^*(B,NT)}{\partial \beta} =& \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \big(NE^{in}(B,NT) + NE^{out}(B,NT)\big)\\ =& \frac{1}{2} \cdot \frac{\partial R^*}{\partial \beta} \cdot \left[\frac{-\alpha(2\delta+1)(\lambda_N-1)(\omega-1)+\omega+2}{4}\right]\end{aligned}$$

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

It is easy to verify that $\overline{\lambda_N} > \underline{\lambda_N}$ 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 4. $\square$

Proof of Proposition 5 Below we prove the two claims outlined in Proposition 5 by showing that their qualitative implications resemble those in the baseline model.

(i) $\frac{\partial P_c(A,NT)}{\partial \beta}$, $\frac{\partial P_c(B,NT)}{\partial \beta}$ and $\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta}$

First, we show that $\frac{\partial R^*}{\partial \beta} > 0$ holds in this extended model.

$$\frac{\partial R^*}{\partial \beta} = \frac{4x\,N(x)}{\begin{bmatrix}\alpha(4\delta^2-1)(\omega-1)+\alpha(\beta-1)^2(\kappa-1)x^2(\omega-1)\\ +x\Big(\alpha\big((\beta-2)\beta+2\big)(\omega-1)+\big(-3(\beta-2)\beta-2\big)\omega+2\Big)-\omega-10\end{bmatrix}^2}.$$

where

$$N(x) = \alpha(\omega-1)\left(2\beta-4\delta^2-1\right)-6\beta\omega+\alpha\left(\beta^2-1\right)(\kappa-1)x^2(\omega-1)$$
$$+x\left(\alpha(\omega-1)(2(\beta-1)\kappa+(\beta-2)\beta)-3\beta^2\omega+2\omega-2\right)+7\omega+10.$$

Here, the denominator is a square, and $4xN(x) > 0$. So as long as the $N(x)$ is positive, we will have $\frac{\partial R^*}{\partial \beta} > 0$.

Note that $N(x)$ is a quadratic function of x and the coefficient of x^2 is $\alpha\left(\beta^2-1\right)(\kappa-1)(\omega-1) < 0$. So the minimum of $N(x)$ is at the endpoints. That is, it is sufficient to show that

$$\min_{x\in[0,\frac{3}{4}]} N(x) = \min\{N(0), N(\frac{3}{4})\} > 0$$

Step 1.1: $N(0)$

$$\begin{aligned} N(0) &= \alpha(\omega-1)\left(2\beta-4\delta^2-1\right)+(7-6\beta)\omega+10 \\ &\geq -2\alpha-7+10 \\ &> 0 \end{aligned}$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\beta \in [0, 1]$ and $\omega \in [-1, 0)$. The last inequality follows $\alpha \in [0, 1]$.

Step 1.2: $N(\frac{3}{4})$

$$\begin{aligned} N(\frac{3}{4}) = &\frac{1}{16}\left(\alpha(\omega-1)\left(\beta(3\beta+8)(3\kappa+1)-64\delta^2-33\kappa-7\right)\right) \\ &+\frac{1}{4}(-3\beta(3\beta+8)\omega+34(\omega+1)). \end{aligned}$$

Since $N(\frac{3}{4})$ is a linear function in ω, we have $\min N(\frac{3}{4}) = \min\{N(\frac{3}{4})|_{\omega=-1}, N(\frac{3}{4})|_{\omega=0}\}$.

$$\begin{aligned} N(\frac{3}{4})|_{\omega=0} &= \frac{1}{16}\alpha\left(-\beta(3\beta+8)(3\kappa+1)+64\delta^2+33\kappa+7\right)+\frac{17}{2} \\ &\geq \frac{1}{16}(-\beta(3\beta+8)+7)+\frac{17}{2} \\ &\geq -\frac{1}{4}+\frac{17}{2} \\ &> 0. \end{aligned}$$

The third last inequality follows $N(\frac{3}{4})|_{\omega=0}$ increases with κ. This is because we have $\frac{\partial N(\frac{3}{4})|_{\omega=0}}{\partial \kappa} = \frac{1}{16}\alpha(33-3\beta(3\beta+8)) > 0$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=-1} &= \frac{1}{8}\left(\alpha\left(-\beta(3\beta+8)(3\kappa+1)+64\delta^2+33\kappa+7\right)+6\beta(3\beta+8)\right) \\
&\geq \frac{1}{8}(-\beta(3\beta+8)+7+18\beta^2+48\beta) \\
&= \frac{1}{8}(15\beta^2+40\beta+7) \\
&> 0.
\end{aligned}
$$

The third last inequality follows $N(\frac{3}{4})|_{\omega=-1}$ increases with κ. This is because we have $\frac{\partial N(\frac{3}{4})|_{\omega=-1}}{\partial\kappa} = \frac{1}{16}\alpha(33-3\beta(3\beta+8)) > 0$. The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\kappa \in [0,1]$ and $\alpha \in [0,1]$ and the last inequality in both expressions follows $\beta \in (0,1)$. Combine these two together, we obtain $N(\frac{3}{4}) \geq \min\{N(\frac{3}{4})|_{\omega=0}, N(\frac{3}{4})|_{\omega=-1}\} > 0$. Since both $N(0) > 0$ and $N(\frac{3}{4}) > 0$, we have $N(x) > 0$ for all $x \in (0, \frac{3}{4})$. Therefore, the numerator $4xN(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial R^*}{\partial\beta} > 0$.

Now that we have established $\frac{\partial R^*}{\partial\beta} > 0$, the rest of the expressions and proof are identical to those of Propositions 1 and 2.

(ii) Show $\frac{\partial[(1-\beta)P_c^*(i,T)]}{\partial\kappa} > 0$ for $i \in \{A, B\}$

The proof involves two steps:

Step 1: show $\frac{\partial[(1-\beta)P_c^*(A,T)]}{\partial\kappa} > 0$.

$$
\frac{\partial[(1-\beta)P_c^*(A,T)]}{\partial\kappa} = \frac{\alpha(\beta-1)^2x(1-\omega)(\beta x+1)\,N_A(x)}{2\begin{bmatrix}\alpha(4\delta^2-1)(\omega-1)+\alpha(\beta-1)^2(\kappa-1)x^2(\omega-1) \\ +x\Big(\alpha\big((\beta-2)\beta+2\big)(\omega-1)+\big(-3(\beta-2)\beta-2\big)\omega+2\Big)-\omega-10\end{bmatrix}^2},
$$

where

$$
N_A(x) = \left(-\alpha\left(4\delta^2-1\right)(\omega-1)+x\left(\alpha(\omega-1)\left(2(\beta-1)^2\delta-1\right)-\omega-2\right)+\omega+10\right).
$$

The denominator is positive, and $\alpha(\beta-1)^2x(1-\omega)(\beta x+1) > 0$ follows from $\beta \in [0,1]$, $x \in (0, \frac{3}{4}]$ and $\alpha \in [0,1]$. Thus, it is sufficient to show that

$$
\min_{x\in(0,\frac{3}{4})} N_A(x) = \min\{N_A(0), N_A(\frac{3}{4})\} > 0
$$

Step 1.1: $N_A(0)$

$$
N_A(0) = -\,\alpha\left(4\delta^2-1\right)(\omega-1)+\omega+10 > 0.
$$

Step 1.2: $N_A(\frac{3}{4})$

$$\begin{aligned} N_A(\frac{3}{4}) &= \frac{1}{4}\left(-\alpha(\omega-1)\left(-6(\beta-1)^2\delta+16\delta^2-1\right)+\omega+34\right) \\ &\geq \frac{1}{4}(\alpha(1-\omega)+\omega+34) \\ &> 0. \end{aligned}$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\beta \in (0,1)$ and $\alpha \in [0,1]$ and the last inequality in both expressions follows $\omega \in [-1,0]$. As a result, we have both $N_A(0) > 0$ and $N_A(\frac{3}{4}) > 0$, Altogether, we obtain $N_A(x) > 0$ for all $x \in (0, \frac{3}{4})$. Therefore, the numerator $\alpha(\beta-1)^2 x(1-\omega)(\beta x+1)N_A(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial[(1-\beta)P_c^*(A,T)]}{\partial\kappa} > 0$.

Step 2: show $\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\kappa} > 0$.

$$\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\kappa} = \frac{\alpha(1-\beta)^2x(1-\omega)(\beta x+1)\,N_B(x)}{2\begin{bmatrix}\alpha(4\delta^2-1)(\omega-1)+\alpha(\beta-1)^2(\kappa-1)x^2(\omega-1) \\ +x\Big(\alpha\big((\beta-2)\beta+2\big)(\omega-1)+\big(-3(\beta-2)\beta-2\big)\omega+2\Big)-\omega-10\end{bmatrix}^2},$$

where

$$N_B(x) = -\left(\alpha\left(4\delta^2-1\right)(\omega-1)+x\left(\alpha(\omega-1)\left(2(\beta-1)^2\delta+1\right)+\omega+2\right)-\omega-10\right).$$

The denominator is always positive, and $\alpha(\beta-1)^2x(1-\omega)(\beta x+1) > 0$ follows from $\beta \in [0,1]$, $x \in (0, \frac{3}{4}]$, and $\alpha \in [0,1]$. It is sufficient to show that

$$\min_{x\in[0,\frac{3}{4}]} N_B(x) = \min\{N_B(0), N_B(\frac{3}{4})\} > 0$$

Step 2.1: $N_B(0)$

$$N_B(0) = -\,\alpha\left(4\delta^2-1\right)(\omega-1)+\omega+10 > 0$$

Step 2.2: $N_B(\frac{3}{4})$

$$\begin{aligned} N_B(\frac{3}{4}) &= \frac{1}{4}\left(-\alpha(\omega-1)\left(6(\beta-1)^2\delta+16\delta^2-1\right)+\omega+34\right) \\ &\geq \frac{1}{4}(\alpha(\omega-1)+\omega+34) \\ &> 0 \end{aligned}$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\beta \in (0,1)$, and $\alpha \in [0,1]$ and the last inequality in both expressions follows $\omega \in [-1, 0)$. As a result, we have both $N_B(0) > 0$ and $N_B(\frac{3}{4}) > 0$, Altogether, we obtain $N_B(x) > 0$ for all $x \in (0, \frac{3}{4}]$. Therefore, the numerator $\alpha(\beta - 1)^2 x(1-\omega)(\beta x+1)N_B(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial[(1-\beta)P_c^*(B,T)]}{\partial\kappa} > 0$.

□

Proof of Proposition 6 Below we prove the two claims outlined in Proposition 6 below, by showing that their qualitative implications resemble those in the baseline model.

(i) $\frac{\partial P_c(A,NT)}{\partial\beta}$, $\frac{\partial P_c(B,NT)}{\partial\beta}$ and $\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial\beta}$

The proof involves two steps.

Step 1: Show $\frac{\partial R^*}{\partial\beta} > 0$ in this extended model

$$\frac{\partial R^*}{\partial\beta} = \frac{4xN(x)}{\left(-\alpha\left(4\delta^2-1\right)(\omega-1)+x((\alpha-3)(\beta-2)\beta\omega+\alpha((\beta-2)\beta+2)-4\omega-2)+\omega+10\right)^2},$$

where

$$\begin{aligned} N(x) =& \alpha\left(-2\beta(\omega+1)-4\delta^2(\omega-1)+3\omega\right)+\alpha+6\beta\omega \\ & -x\left((\alpha-3)\beta^2\omega+\alpha\left(\beta^2-2\right)+4\omega+2\right)-5\omega+10. \end{aligned}$$

The denominator is always positive and $4x > 0$ follows from $x \in (0, \frac{3}{4}]$. It is sufficient to show that

$$\min_{x\in[0,\frac{3}{4}]} N(x) = \min\{N(0), N(\frac{3}{4})\} > 0.$$

Step 1.1: $N(0)$

$$\begin{aligned} N(0) &= \alpha\left(-2\beta(\omega+1)-4\delta^2(\omega-1)+3\omega+1\right)+(6\beta-5)\omega+10 \\ &\geq -2-1+10 \\ &> 0. \end{aligned}$$

The second inequality follows from $\omega \in [-1, 0)$, $\alpha \in [0,1]$, and $N(0)$ decreases with β for $\beta \in [0,1]$.

Step 1.2: $N(\frac{3}{4})$

$$\begin{aligned} N(\frac{3}{4}) =& -\frac{1}{4}\alpha\left(3\beta^2(\omega+1)+8\beta(\omega+1)+2\left(8\delta^2(\omega-1)-6\omega-5\right)\right) \\ & +\left(\frac{9\beta^2}{4}+6\beta-8\right)\omega+\frac{17}{2} \end{aligned}$$

Note that $N(\frac{3}{4})$ is a linear function in ω. Hence, $\min_{x\in[0,\frac{3}{4}]} N(\frac{3}{4}) = \min\{N(\frac{3}{4})|_{\omega=-1}, N(\frac{3}{4})|_{\omega=0}\}$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=0} &= \frac{1}{4}\left(\alpha\left(-3\beta^2-8\beta+16\delta^2+10\right)+34\right)\\
&\geq \frac{1}{4}(-3-8+10)+\frac{17}{2}\\
&= \frac{33}{4}\\
&> 0.
\end{aligned}
$$

The third last inequality follows $N(\frac{3}{4})|_{\omega=0}$ decreases with β for $\beta \in [0,1]$.

$$
\begin{aligned}
N(\frac{3}{4})|_{\omega=-1} &= \frac{1}{4}\left(\alpha\left(32\delta^2-2\right)-3\beta(3\beta+8)+66\right)\\
&\geq \frac{1}{4}(-2-3(3+8)+66)\\
&= \frac{31}{4}\\
&> 0.
\end{aligned}
$$

The second last inequality in both expressions follows from $\delta \in [0, \frac{1}{2}]$, $\kappa \in [0,1]$ and $\alpha \in [0,1]$ and the last inequality in both expressions follows $\beta \in (0,1)$. Combine these two together, we obtain $N(\frac{3}{4}) \geq \min\{N(\frac{3}{4})|_{\omega=0}, N(\frac{3}{4})|_{\omega=-1}\} > 0$.Since both $N(0) > 0$ and $N(\frac{3}{4}) > 0$, we have $N(x) > 0$ for all $x \in (0, \frac{3}{4})$. Therefore, the numerator $4xN(x) > 0$, and the denominator is positive. Hence, we have $\frac{\partial R^*}{\partial \beta} > 0$.

Step 2: $\frac{\partial P_c(i,NT)}{\partial \beta}$ and $\frac{\partial \mathbb{E}[U^r(A)-U^r(B)]}{\partial \beta}$

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. Therefore, the engagement received by the non-toxic creators do not change. With $\frac{\partial R^*}{\partial \beta} > 0$, the analysis and expressions are identical to the proofs of Propositions 1 and 2.

(ii) Surviving toxic content is higher.

The key feature of this extended model is that toxic users value negative engagement, which increases their total received engagement, $[NE^{in}(i,T) + NE^{out}(i,T)]$, relative to our baseline model. Recall that the creation probability for a toxic creator is given by:

$$
P_c^*(i,T) = \frac{1}{2} + \frac{1}{2}\cdot(1-\beta)\cdot R^*[NE^{in}(i,T) + NE^{out}(i,T)].
$$

The total effect of change in this engagement term on content creation can be decomposed into two components: direct and indirect effect. The direct effect is $\frac{1}{2}(1-\beta)R^*$ and the indirect (equilibrium) effect is through its effect on expected reach, given by $-\lambda(i,T)\cdot\frac{1}{2}(1-\beta)R^*$. The net effect is the sum of these two components. Since $\lambda(i, T) < 1$ and we have $(1 - \beta)R^* > 0$, the direct effect dominates and thus the net effect is positive. This implies that the higher engagement from valuing negative content leads to an overall increase in content creation for toxic creators.

$\square$

A.3 Alternative Reach Specifications

Our main analysis models the crowding effect on content reach ($R(\cdot)$) through the linear approximation in Equation (3), which captures how total content supply reduces the content reach of any given post. To verify that our qualitative results are not an artifact of this linear form, we conducted numerical simulations under several alternative specifications of $R(\cdot)$ that preserve the same monotone "crowding-out" property but differ in curvature and boundary behavior.

Setup. We re-solve the equilibrium numerically for each functional form using the same primitives as in the main text. Specifically, we consider four common variants:

$$\text{(i)} \quad R^{\text{linear}} = 1 - V, \qquad \text{(ii)} \quad R^{\text{inverse}} = \frac{1}{1+V},$$
$$\text{(iii)} \quad R^{\text{exponential}} = e^{-V}, \qquad \text{(iv)} \quad R^{\text{logistic}} = \frac{1}{1+\exp(V-\frac{1}{2})},$$

where V denotes the total volume of surviving content, as in the baseline model. All forms above ensure $R \in (0,1]$ and $\partial R/\partial V < 0$. The linear form used in the main text corresponds to case (i). The inverse and exponential forms (cases ii and iii) reduce locally to the linear case for small V.

Implementation. For each $R(\cdot)$, we solve the fixed-point system described in Section 2 using the bounded solver from MATLAB. Comparative statics with respect to β, α, and δ are computed via centered finite differences, as in the main analysis. Figure 5 summarizes results for $\partial P_c^*(i,t)/\partial\beta$ across the parameter grid. All simulations use identical primitives and initialization to ensure comparability.[9]

Findings. Across all specifications, we obtain the same qualitative pattern as in Figure 2. That is, increasing β (a stricter moderation environment) can polarize, universally empower or suppress majority and minority group non-toxic creators, depending on the levels of α and δ. Figure 5 shows that the sign structure of the cross-partials $\partial P_c^*(i,NT)/\partial\beta$ is preserved across all functional forms of $R(\cdot)$, confirming that our main insights are robust to the various representations of crowding effect. The linear specification therefore offers a most tractable and representative benchmark.

[9]The MATLAB scripts and parameter settings are available upon request.

Majority Creator A: $\partial P_c^*(A, NT)/\partial\beta$ **Minority Creator B:** $\partial P_c^*(B, NT)/\partial\beta$

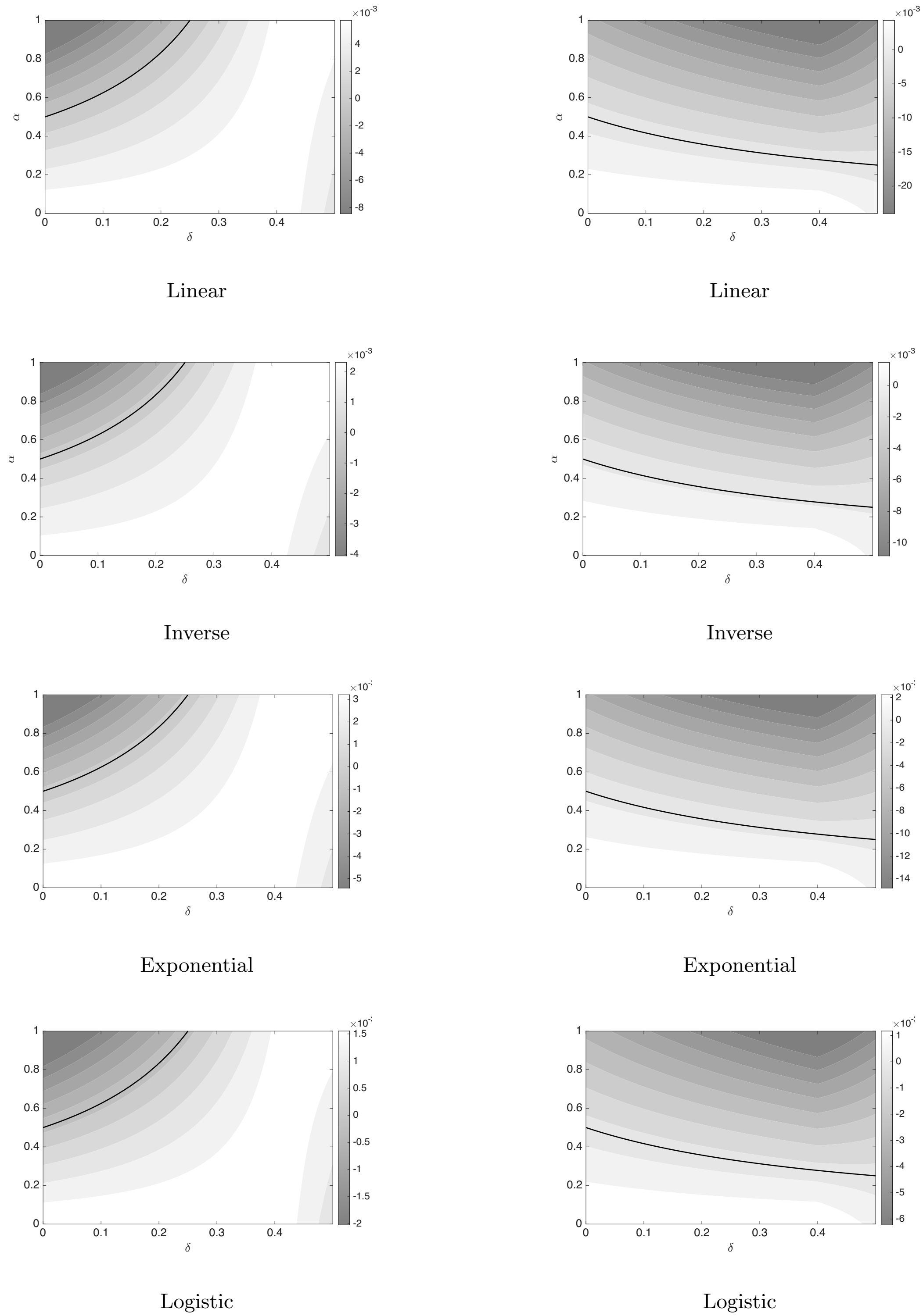


Figure 5: Comparative statics of $\partial P_c^*(i, NT)/\partial\beta$ for non-toxic users from ideological majority group A (left column) and minority group B (right column), under four reach functions: linear, inverse, exponential, and logistic. Each panel includes the zero–contour (black solid line). Parameters: $\omega = -1$, $\beta = 0.5$, $x = 0.2$, $\tau_A = 0.5$.