Spiral of Silence:

How Neutral Moderation Polarizes Content Creation

Ying Bao* Jessie Liu†

* University of Illinois at Urbana-Champaign
† Johns Hopkins University, Carey Business School

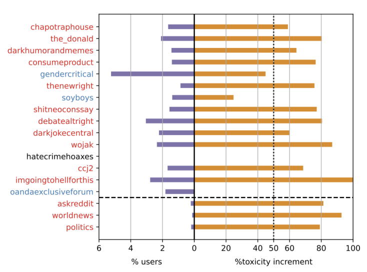
Online Toxicity – A Growing Concern

- About 4 in 10 Americans have faced online harassment;
- 71% support tighter platform rules on toxic content (Pew Research Center, 2021).

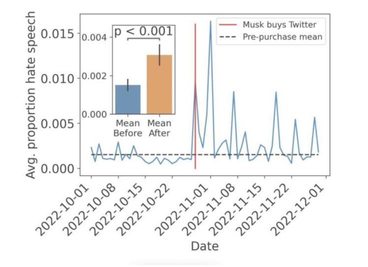
Online Toxicity – A Growing Concern

- About 4 in 10 Americans have faced online harassment;
- 71% support tighter platform rules on toxic content (Pew Research Center, 2021).

A Puzzle: Opposite Moderation Tools, Same Outcome



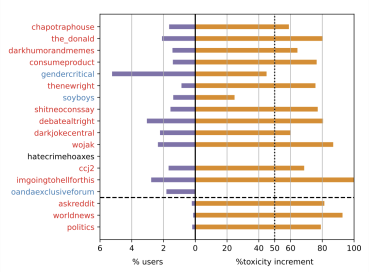
Cima et al., arXiv, 2024



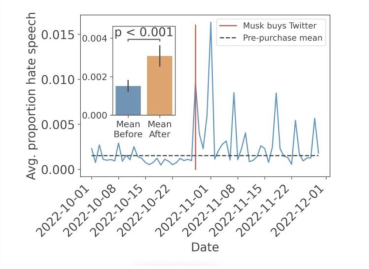
Burghardt et al., Phys.org, 2023

- When Reddit **banned** toxic communities, some toxic users become **more** active whereas the less-toxic ones, posted **less**.

A Puzzle: Opposite Moderation Tools, Same Outcome



Cima et al., arXiv, 2024



Burghardt et al., Phys.org, 2023

- When Reddit **banned** toxic communities, some toxic users become **more** active whereas the less-toxic ones, posted **less**.
- When Twitter **relaxed** its rules, hate speech **doubled** too.

The Question

If both moderation and deregulation increases toxicity, then *what*, is being moderated?

The Question

If both moderation and deregulation increases toxicity, then *what*, is being moderated?

- Most existing studies focus on **demand**-side trade-off:
  - **freedom** of expression vs. **harms** from online content (e.g. Liu et al, 2022; Beknazar-Yuzbashev et al., 2024)

The Question

If both moderation and deregulation increases toxicity, then *what*, is being moderated?

- Most existing studies focus on **demand**-side trade-off:
  - **freedom** of expression vs. **harms** from online content (e.g. Liu et al, 2022; Beknazar-Yuzbashev et al., 2024)
- But moderation changes not only "who sees what", but also *who* finds it worthwhile to **create**.

Our paper: From Reaction to Anticipation

- On social media, every user is **both** a consumer and a (potential) creator.
  - Creation *anticipates* engagement.
  - Consumption *reacts* to what survives moderation.

Our paper: From Reaction to Anticipation

- On social media, every user is **both** a consumer and a (potential) creator.
  - Creation *anticipates* engagement.
  - Consumption *reacts* to what survives moderation.
- Together they form a **feedback loop:**

$$\begin{tikzcd} \begin{array}{c}\text{Creation}\\ \text{(anticipation)}\end{array} \arrow[r] & \text{Content Pool} \arrow[r] & \begin{array}{c}\text{Consumption}\\ \text{(reaction)}\end{array} \arrow[r] & \begin{array}{c}\text{Engagement}\\ (+\ /\ -)\end{array} \arrow[lll, bend left] \end{tikzcd}$$

Our paper: From Reaction to Anticipation

- On social media, every user is **both** a consumer and a (potential) creator.
  - Creation *anticipates* engagement.
  - Consumption *reacts* to what survives moderation.
- Together they form a **feedback loop:**

$$\begin{tikzcd} \begin{array}{c}\text{Creation}\\ \text{(anticipation)}\end{array} \arrow[r] & \text{Content Pool} \arrow[r] & \begin{array}{c}\text{Consumption}\\ \text{(reaction)}\end{array} \arrow[r] & \begin{array}{c}\text{Engagement}\\ (+ \,/\, -)\end{array} \arrow[lll, bend left] \end{tikzcd}$$

- <u>Our focus</u>: Moderation shapes not only what is removed, but what is never created.

▸ Lit Review

A New Kind of Market Failure

- Content has two dimensions:
  - **Vertical:** Toxicity — *how* it is said.
  - **Horizontal:** Ideology — *what* is said, and to whom.

A New Kind of Market Failure

- Content has two dimensions:
  - **Vertical:** Toxicity — *how* it is said.
  - **Horizontal:** Ideology — *what* is said, and to whom.
- Moderation tools (e.g., Perspective API) are trained on data to target the vertical, *not* the horizontal.

A New Kind of Market Failure

- Content has two dimensions:
  - **Vertical:** Toxicity — *how* it is said.
  - **Horizontal:** Ideology — *what* is said, and to whom.
- Moderation tools (e.g., Perspective API) are trained on data to target the vertical, *not* the horizontal.
- This creates an unpriced **relational externality**:
  - Each created (or removed) post changes the relational network of in-group favoritism vs. out-group animosity, and with it, other creators' incentives.

A New Kind of Market Failure

- Content has two dimensions:
  - **Vertical:** Toxicity — *how* it is said.
  - **Horizontal:** Ideology — *what* is said, and to whom.
- Moderation tools (e.g., Perspective API) are trained on data to target the vertical, *not* the horizontal.
- This creates an unpriced **relational externality**:
  - Each created (or removed) post changes the relational network of in-group favoritism vs. out-group animosity, and with it, other creators' incentives.
- Even **neutral** policy can **polarize** content creation: a structural *market failure* of social networks.

Roadmap

- Main Model.
- Extensions:
  - **Personalization**: ideology-based curation
  - **Neutrals**: indifferent to partisan ideology
  - **Targeted Toxicity**: toxicity itself becomes relational
- Conclusion and Implications

The Model

- A mass 1 of users differ in two dimensions $(i, t) \in \{A, B\} \times \{T, NT\}$:
  - ideology type $i \in \{A, B\}$ reflects their ideology.
  - toxicity type $t \in \{T, NT\}$ reflects their propensity to create toxic content.
- All users have a (potential) dual role as:
  - **readers** who read content created by others,
  - **creators** who post content that are read by others.

Content Consumers: Utility

A reader of type (i, t) derives utility U from consuming content created by a type (i', t') user:

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

where

Content Consumers: Utility

A reader of type (i, t) derives utility U from consuming content created by a type (i', t') user:

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

where $\alpha \in (0, 1)$ calibrates the *relative* importance of

- **horizontal** preference: $H(i, i') = \begin{cases} 0 & \text{if } i = i' \\ -1 & \text{if } i \neq i' \end{cases}$

Content Consumers: Utility

A reader of type (i, t) derives utility U from consuming content created by a type (i', t') user:

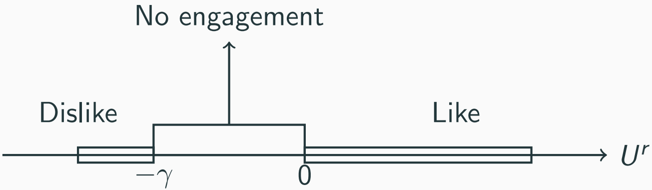
$$U^r(i, i't') = \underbrace{\alpha \cdot H(i, i')}_{\text{horizontal (ideology-based)}} + \underbrace{(1-\alpha) \cdot V(t')}_{\text{vertical (toxicity-based)}} + \varepsilon_r$$

where $\alpha \in (0, 1)$ calibrates the *relative* importance of

- **horizontal** preference: $H(i, i') = \begin{cases} 0 & \text{if } i = i' \\ -1 & \text{if } i \neq i' \end{cases}$
- **vertical** preference: $V(t') = \begin{cases} 0 & \text{if } t' = NT \\ -1 & \text{if } t' = T \end{cases}$

▸ Back

Content Consumers: Engagement



- A reader i will **like** a post if $U^r(\cdot) \geq 0$.
- A reader i will **dislike** a post if $U^r(\cdot) < -\gamma$, where $\gamma \in [0, 1]$ calibrates the cost of dislike relative to that of likes.
  - Consistent with research (e.g. Stsiampkouskaya et al., 2023), users are more inclined to "like" rather than "dislike"
  - W.L.O.G. we normalize $\gamma = \frac{1}{2}$.

Engagement Probability

The hallmarks of **affective polarization** (α) (Iyengar, 2012):

- In-group *favoritism*: Readers react more positively to (even toxic) content from creators with the **same** ideology.
- Out-group *animosity*: Readers react more negatively to (even non-toxic) content from creators with **opposite** ideology.

Content Creators

- Given their type (i, t), each user decides whether to create a post $a \in \{0, 1\}$. Utility from creation $(a = 1)$ is given by

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

Content Creators

- Given their type (i, t), each user decides whether to create a post $a \in \{0, 1\}$. Utility from creation $(a = 1)$ is given by
  $$U^c(i,t) = \underbrace{S(t)}_{\text{survival rate}} \cdot \underbrace{R(V)}_{\text{content reach}} \cdot \underbrace{\left(NE^{in}(i,t) + NE^{out}(i,t)\right)}_{\text{engagement}} \underbrace{+\varepsilon_c}_{\text{intrinsic utility}}$$
  where
  $$NE^{in}(i,t) = \underbrace{\lambda_i}_{\text{in-group share}} \cdot [\underbrace{P_l(i,it)}_{\text{prob. of like}} + \omega \underbrace{P_{dl}(i,it)}_{\text{prob. of dislike}}],$$
  $$NE^{out}(i,t) = \underbrace{\lambda_{-i}}_{\text{out-group share}} \cdot [P_l(-i,it) + \omega P_{dl}(-i,it)],$$
- $\omega \in [-1, 0)$: the relative *weight* of negative over positive engagement on content creation (we relax it later).

Extension: $\omega > 0$

Content Moderation and Reach

For analytical tractability, we model moderation intensity via a **survival** rate:

$$S(t) = 1 - \underbrace{\beta}_{\text{\% toxic posts removed}} \cdot \mathbf{1}[t = T];$$

and content **reach** (from the creator's perspective) as:

$$R(V) = 1 - V; \quad V = \sum_{i,t} \underbrace{\lambda(i,t)}_{\text{share}} \cdot S(t) \cdot \underbrace{P_c(i,t)}_{\text{creation prob.}}$$

where V is total volume of (surviving) content

Content Moderation and Reach

For analytical tractability, we model moderation intensity via a **survival** rate:

$$S(t) = 1 - \underbrace{\beta}_{\text{\% toxic posts removed}} \cdot \mathbf{1}[t = T];$$

and content **reach** (from the creator's perspective) as:

$$R(V) = 1 - V; \quad V = \sum_{i,t} \underbrace{\lambda(i,t)}_{\text{share}} \cdot S(t) \cdot \underbrace{P_c(i,t)}_{\text{creation prob.}}$$

where V is total volume of (surviving) content

- **Crowding-out effect** $(R'(\cdot) < 0)$: More content from others **reduces** the chance any *single* post is seen — creators compete for limited user attention.

Equilibrium Concept

Rational Expectation Equilibrium: In equilibrium, the expected share of each user type that creates content *equals* the probability of creation implied by their utility.

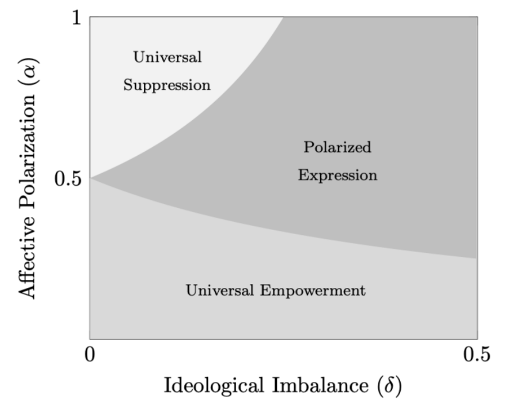
$$P_c(i,t) = Pr[U^c(i,t) > 0],$$
$$\forall i \in \{A,B\}, t \in \{T,NT\}$$

A unique solution exists by Brouwer fixed-point theorem.

Core Result: Spiral of Silence as Equilibrium

Proposition 1(c). Moderation Polarizes Creation: $\forall \beta \in [0,1]$,

$$\alpha \in (\underline{\alpha}(\delta), \overline{\alpha}(\delta)) \Rightarrow \frac{\partial P_c^*(A, NT)}{\partial \beta} > 0 > \frac{\partial P_c^*(B, NT)}{\partial \beta},$$



Core Result: Spiral of Silence as Equilibrium

- Each takedown removes a toxic post, but also **reweights** *expectations* of in- vs. out-group exposure.
- This reweighting amplifies the *impact* of **asymmetry**:
  - **Ideological imbalance** (δ): unequal group sizes creates asymmetry in influence.
  - **Affective polarization** (α): an emotional divide deepens in-group favoritism and out-group animosity.

Core Result: Spiral of Silence as Equilibrium

- Each takedown removes a toxic post, but also **reweights** *expectations* of in- vs. out-group exposure.
- This reweighting amplifies the *impact* of **asymmetry**:
  - **Ideological imbalance** (δ): unequal group sizes creates asymmetry in influence.
  - **Affective polarization** (α): an emotional divide deepens in-group favoritism and out-group animosity.
- Moderation triggers a feedback loop:
  - The **majority**, sustained by in-group **favoritism**, creates more.
  - The **minority**, discouraged by out-group **animosity**, creates less.

Key Implication: The Appearance of Bias

- Critics call it “censorship”; empirical studies call it “toxicity asymmetry” (Haimson et al., 2021).
- Our model shows a <u>third</u> logic:
  what looks like bias can endogenously *arise* as equilibrium **self-censorship**.
- The “spiral of silence” (Noelle-Neumann, 1974) is thus not enforced; but **induced** by moderation.

Average Wins Can Hide widening Gaps

- **Proposition 2.** Moderation redistributes reader welfare:
  For all $\alpha > 0, \delta > 0$, $\beta > 0$, we have:

$$\frac{dE[\Delta U]}{d\beta} = \frac{dE[U^r(A) - U^r(B)]}{d\beta} > 0$$

Average Wins Can Hide widening Gaps

- **Proposition 2.** Moderation redistributes reader welfare:
  For all $\alpha > 0, \delta > 0$, $\beta > 0$, we have:

$$\frac{dE[\Delta U]}{d\beta} = \frac{dE[U^r(A) - U^r(B)]}{d\beta} > 0$$

- Both ideological groups benefit from reduced exposure to toxic content (a universal gain);
- However, the composition of what remains tilts further toward content of ideological majority (a polarizing effect).

Average Wins Can Hide widening Gaps

- **Proposition 2.** Moderation redistributes reader welfare:
  For all $\alpha > 0, \delta > 0$, $\beta > 0$, we have:

$$\frac{dE[\Delta U]}{d\beta} = \frac{dE[U^r(A) - U^r(B)]}{d\beta} > 0$$

- Both ideological groups benefit from reduced exposure to toxic content (a universal gain);
- However, the composition of what remains tilts further toward content of ideological majority (a polarizing effect).
- Relational Externality: majority content gain higher reach without internalizing its **negative spillover** on the minority.

Extension 1: Content Personalization

We modify the effective content reach of in-group vs. out-group as:

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

where

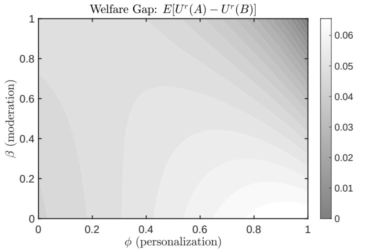
$$R^{in}(i,t) = R(\cdot), \qquad R^{out}(i,t) = (1 - \phi)\, R(\cdot);$$

- $\phi \in [0,1]$ captures how much personalization filters out cross-group exposure.

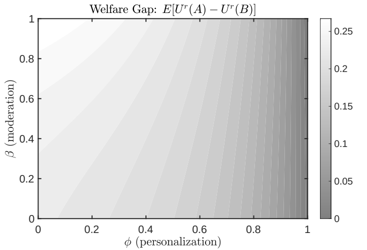
Personalization vs. Moderation

- Personalization is *not* a perfect complement for moderation:
  - **Conditional** relief on **non-toxic** creators (more in-group support, but also narrowed cross-group reach)
  - **Revive** minority **toxic** creation (via sympathetic in-group).
- Personalization could create a **subtler form of polarization** beyond echo chambers:
  - protects minority creation yet reshapes the content mix towards more toxicity;
  - even exacerbates "spiral of silence" at low α.

Personalization vs. Moderation



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



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

- Polarized spaces (e.g., X/Twitter): moderate personalization (ϕ) + strong moderation (β)
- Knowledge-based (e.g., Reddit/Wikipedia): too much personalization is counterproductive

Extension 2: Ideological Neutrals

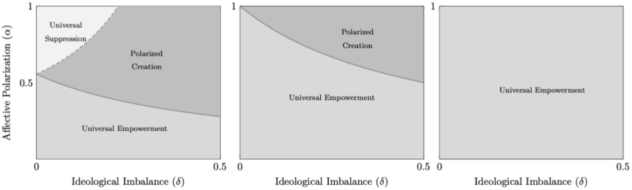
- Neutrals ($i = N$) do not care about ideology, only toxicity;
- Partisans ($i = A, B$) are more tolerable of neutral content:

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

- Neutrals **dilute** the (affective) tension between partisans, but their creation also **crowds out** reach of partisan content.

Extension 2: Ideological Neutrals

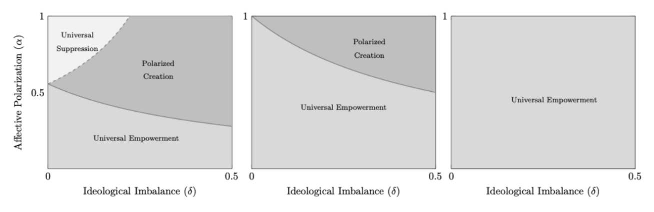
- When the share of neutral is **moderate**, the crowd-out effect dominates the animosity dilution effect, recreating "spiral of silence" (among partisans).



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

Extension 2: Ideological Neutrals

- When the share of neutral is **moderate**, the crowd-out effect dominates the animosity dilution effect, recreating "spiral of silence" (among partisans).



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

- When neutrals form a **small minority**, stricter moderation can **deepen their own silence** despite declining toxicity on the platform (a parallel of "spiral of silence" among neutrals).

Extension 3: Targeted Toxicity Homophily

When toxicity itself becomes **relational**:

- Toxic readers value toxicity - when it targets the out-group.

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

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

  where $\kappa \in (0, 1)$ captures the strength of such toxicity homophily.

- All qualitative results remain the same; but (surviving) toxic posts increases with κ.

Summary

- "Neutral" moderation is **not** neutral in outcome.
  - Moderation changes not just what gets removed, but
  - the relational environment that determines what gets *created* in equilibrium.
- Moderation generates content-level **externality**:
  - Affective polarization and ideological imbalance jointly create the asymmetry of emotional divide
- Regulating one dimension can distort the **supply** of another: toxicity (moderation) $\leftrightarrow$ ideology (personalization).

"Scientific inquiry begins and ends with problems."

– Karl Popper, Conjectures and Refutations (1963)

Our model makes no claim to completeness, but it could be *useful*:

"Scientific inquiry begins and ends with problems."

– Karl Popper, Conjectures and Refutations (1963)

Our model makes no claim to completeness, but it could be *useful*:

- **For policy:** suggests diagnose before intervention
  - Mislabeling disagreement as toxicity (because it's unpopular among the majority) could institutionalize "spiral of silence."
- **For marketing:** reframes **affective polarization**, a core concept in political science, as a *structural feature* of social networks, which extends beyond political ideology per se (Lerman et al,, 2024).