Contestability and the Optimal Regulation of Social Media Platforms

Martino Banchio Francesco Decarolis Carl-Christian Groh
Rafael Jiménez-Durán Miguel Risco

Digital Competition Conference 2026 — Knight-Georgetown Institute

February 5, 2026

Research Question

What is the optimal regulatory approach
for social media platform markets?

The core tension:

- Platforms profit from **engagement**
  $\rightarrow$ incentives to display harmful but engaging content
- Harmful content **hurts users** (mental health, polarization)
- Many users **don't internalize** these harms when choosing platforms

Standard policy: Reduce dominant platforms' market power to improve contestability

Our finding: When user sophistication is low, this might not be enough

Model: Key Ingredients

Players:

- **Incumbent** (I): Dominant platform with competitive advantage
- **Entrant** (E): Challenger platform
- **Users**: Unit mass, choose which platform to join

Platform choice and objective:

- Each platform $p \in \{I, E\}$ chooses harmful content share $h_p \in [0, 1]$
- Platform revenue $\propto$ engagement of users who join: $\Pi_p = \int_{\{i:j_i=p\}} \pi\left(e_p^*(h_p)\right) di$

User heterogeneity:

- A share ρ is **rational**: internalize harm when choosing platforms
- A share $1 - \rho$ is **naive**: neglect the adverse effects of harmful content

Pew 2025: Many users are aware of harmful effects, but believe do not affect them personally

Model: Key Assumptions

1. **Incumbent advantage:** For all $h \in [0,1]$

$$U_I^r(h) > U_E^r(h) \quad \text{and} \quad U_I^n(h) > U_E^n(h)$$

(Better data, larger network, superior algorithm)

2. **Exposure to harmful content decreases true utility:** $U_p^r(h)$ strictly decreasing in h

$$U_p^r(1) < 0 < U_p^r(0)$$

3. **Naive users neglect the cost of harmful content:**
Perceived utility $U_p^n(h)$ responds to engagement but omits the harm cost
$\rightarrow$ $U_p^n(h)$ weakly increasing in h (via engagement)

4. **Harmful content is more engaging:** $e_p^*(h)$ strictly increasing in h

Timing: First, platforms choose h_p simultaneously. Then, users observe and choose platform.

Equilibrium Regimes

Market Dominance

When: ρ high, advantage large

Mechanism: Incumbent retains rationals by limiting harm

Equilibrium: $h_I^* = \check{h}_I, \quad h_E^* = 0$

Allocation:
All users on incumbent

Highest welfare

Mixed Strategies

When: ρ intermediate

Mechanism: Trade-off: cater to rationals vs. farm naives

Equilibrium: Both platforms randomize over h

Allocation:
Users split across platforms

Intermediate welfare

Naivety-Focused

When: ρ low

Mechanism: Incumbent abandons rationals, farms naives

Equilibrium: $h_I^* = 1, \quad h_E^* = \tilde{h}_E$

Allocation:
Rationals on E; naives on I

Lowest welfare

Key insight: When ρ is low, the incumbent prefers to maximize engagement from naives rather than compete for rationals

Result 1: User Migration Can Hurt

Proposition 1

User welfare is **strictly higher** in any equilibrium in which all users join the incumbent than in any equilibrium in which some users join the entrant.

Intuition:

- **Market dominance:** Incumbent must offer positive utility to retain rational users
  $\rightarrow$ disciplines harmful content
- **User split:** Platform with naive users sets $h = 1$ to maximize engagement
  $\rightarrow$ Rational users get negative utility on such a platform
  $\rightarrow$ Competing platform raises h_E until $U_E^r = 0$ (participation constraint binds)
  $\rightarrow$ Rational users get zero utility; naive users get negative (true) utility

Implication: Policies that induce migration away from incumbent may have non-monotonic effects on welfare

Result 2: Contestability Has Limits When ρ Is Low

Proposition 3

If $\rho < \underline{\rho}$, reducing the incumbent's competitive advantage **cannot benefit users**.

Why? When ρ is low:

- Incumbent *always* sets $h_I = 1$ to maximize engagement from naive users
- Rational users join the entrant, but obtain **zero utility** (entrant raises h_E until participation constraint binds)
- Better entrant technology $\rightarrow$ entrant can raise h_E further while retaining users
- Worse incumbent technology $\rightarrow$ naive users' utility decreases

Implication: When user sophistication is low, portability or interoperability measures would not improve user welfare

Three Regulatory Instruments

1. **Measures that reduce incumbent's advantage**
   - *Interoperability:* DMA Art. 7 (currently scoped to NIICS)
   - *Switching/portability:* GDPR Art. 20; DMA Art. 6(9)

   $\rightarrow$ Reduce the gap $U_I^r(h) - U_E^r(h)$

   $\rightarrow$ When ρ is low: does not benefit users; at moderate ρ: can backfire
2. **Sophistication measures** *(California AB 56, DSA Art. 27)*
   - Warning labels, algorithmic transparency, digital literacy programs
   - Increases $\rho \rightarrow$ reduces harmful content in equilibrium
3. **Content moderation obligations** *(DSA Art. 34–35)*
   - Systemic risk assessment, risk mitigation measures for VLOPs
   - Decreases *effective* harmful content (via mandated risk mitigation) **regardless of** ρ

   $\rightarrow$ truncates the "race to harmful content"

Key: These instruments can work as complements

Two Illustrative Cases

Case A: Instagram vs. TikTok

- Young user demographic suggests relatively low ρ
- Intense competition for engagement; TikTok has a better-performing algorithm, Instagram has a larger user base
- Model predicts: near the mixed-strategy or naivety-focused region
- Implication: contestability measures alone may not improve outcomes

Case B: $\mathbb{X}$ vs. Threads/BlueSky

- Post-acquisition, $\mathbb{X}$ shifted toward harmful yet engaging content
- Sophisticated users migrated to Threads/BlueSky
- Consistent with a transition toward the naivety-focused equilibrium
- Entrants then set h_E so that rational users' participation constraint binds

Both cases illustrate: when ρ is low, competition alone does not discipline harmful content

Takeaways

1. **More competition $\neq$ always better outcomes**

   When user sophistication is low, migration away from dominant platforms can reduce welfare

2. **User sophistication can be the binding constraint**

   When ρ is low, standard competition measures might not improve user welfare
   $\rightarrow$ Awareness campaigns and transparency may need to come first

3. **Combining instruments**

   Content moderation obligations can make contestability measures effective
   Sophistication measures can make competition work as intended

Thank you!

Backup Slides

Backup: User Welfare Definition

User welfare = Expected utility of users

- Naive users' *true* utility is $U_p^r(h_p)$, not perceived utility $U_p^n(h_p)$
- All users on platform p have same engagement level $e_p^*(h_p)$
- Welfare accounts for both rational and naive users

User-optimal outcome:

$$h_I^* = 0 \quad \text{and all users join incumbent}$$

This emerges only when incumbent has **no** competitive advantage and ρ is high

Backup: Naive Users—Formal Specification

True utility (for any user on platform p):

$$U_p^r(h_p, e_i) = \big(\eta_p h_p + \theta_p(1 - h_p)\big)e_i + (1 - h_p) - \delta h_p - \gamma(e_i)^2$$

Perceived utility (for naive users):

$$U_p^n(h_p, e_i) = \big(\eta_p h_p + \theta_p(1 - h_p)\big)e_i + (1 - h_p) - \gamma(e_i)^2$$

Key distinction:

- Naive users respond to engagement benefit but **omit the harm cost**
- Since $e_p^*(h)$ is increasing in h, perceived utility $U_p^n(h)$ is increasing in h
- This is *not* "liking harm"—it is responding to engagement while neglecting harm
- Consistent with optimism bias: "I know social media is harmful, but not to me"

Backup: Equilibrium Thresholds

Key objects:

- $\tilde{h}_p$: Harmful content share at which a user obtains zero utility on platform p

$$U_p^r(\tilde{h}_p) = 0$$

- $\check{h}_I$: Harmful content share at which rational users are indifferent between joining the incumbent and the entrant (if entrant displays no harmful content)

$$U_E^r(0) = U_I^r(\check{h}_I)$$

- $\underline{\rho}$: Threshold below which naivety-focused equilibrium is unique

$$(1 - \underline{\rho})\pi_I^n(e_I^*(1)) = \underline{\rho}\pi_I^r(e_I^*(\tilde{h}_I)) + (1 - \underline{\rho})\pi_I^n(e_I^*(\tilde{h}_I))$$

Backup: Content Moderation Proposition

Proposition (Content Moderation)

Fix a moderation standard $\bar{h} \in (0, 1]$ such that $h_p \leq \bar{h}$.

(i) If $\bar{h} \leq \check{h}_I$: all users join incumbent, $h_I^* = \bar{h}$

(ii) Stricter standards expand the parameter region where market dominance is sustainable

(iii) Content moderation **complements** contestability measures

Intuition:

- Standards truncate the "race to harmful content"
- Make contestability measures "safer" to implement
- Even if entrant improves, standards prevent transition to mixed-strategy equilibrium

Backup: Robustness

Extensions in online appendix:

- **Multi-homing:** Results extend; multi-homers choose zero engagement on one platform
- **Network effects:** Key predictions unchanged
- **Alternative naivety:** Users underestimate harmful content share by factor $\alpha < 1$
  $\rightarrow$ Only market dominance equilibrium in pure strategies
- **Captive users:** Some users "locked in" to incumbent
  $\rightarrow$ Reducing lock-in can *increase* incumbent's market share
- **Personalized content:** Results go through under third-degree discrimination