# Contestability and the Optimal Regulation of Social Media Platforms*

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

October 6, 2025

**Abstract**

We study the optimal regulatory approach for social media markets. We consider a model of platform competition between an entrant and an incumbent platform. Any platform has incentives to display harmful content to its users because this maximizes user engagement, thereby generating more revenue for the platform. Some users are naive: When choosing which platform to join, naive users neglect the detrimental effects of being exposed to harmful content. We show that 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. If the share of naive users is high, reducing the incumbent's competitive advantage cannot benefit users. Reductions of the share of naive users can benefit all users, but may raise the incumbent's market share. The user-optimal outcome emerges if the incumbent has no competitive advantage and the share of naive users is small. This is because every platform always has incentives to display less harmful content than its rival under these conditions.

**Keywords:** Contestability, social media platforms, bounded rationality, welfare

**JEL Codes:** D18, D21, D63, L51

*We would like to thank Luca Braghieri, Christoph Carnehl, Francesc Dilmé, Jan Eeckhout, Chiara Fumagalli, Hans-Peter Grüner, Nenad Kos, Stephan Lauermann, Nicola Limodio, Benny Moldovanu, Volker Nocke, Marco Ottaviani, Fausto Panuzzi, Nicolas Schutz, Fernando Vega-Redondo, Ernst-Ludwig von Thadden, Jonas von Wangenheim, and seminar audiences at Bocconi, Bonn and Valencia for insightful comments. Carl-Christian Groh gratefully acknowledges support from the Deutsche Forschungsgemeinschaft (German Research Foundation) through CRC TR 224. Francesco Decarolis and Miguel Risco gratefully acknowledge support from the ERC Consolidator Grant "CoDiM" (GA No: 101002867).

†Bocconi University

‡University of Bonn

# 1 Introduction

Social media platforms are at the center of public and regulatory attention nowadays. This reflects the building amount of evidence that the presence of these platforms contributes to major societal issues: For example, it is well-documented that social media usage has massive detrimental effects on the mental health of users (Sadagheyani and Tatari, 2021; Braghieri et al., 2022). Moreover, the algorithms implemented by 𝕏 and TikTok have been identified to shift political opinions towards extreme positions (Guardian, 2021; DW, 2025). One key issue that underlies the adverse effects of social media is the platforms' core business model, which introduces a tension between user welfare and platform profits: Social media platforms have incentives to display maximally engaging content to their users, i.e., content that maximizes the time users spend on the platform, even if this causes their users harm (Rosenquist et al., 2021; Scott Morton and Dinielli, 2022).

In this paper, we study the optimal regulatory approach for social media platform markets, taking explicit account of the well-documented fact that a majority of social media users do not internalize the detrimental effects of being exposed to harmful content when choosing which platforms to join (Hoong, 2021; Allcott et al., 2022).[1] Existing regulatory proposals for social media markets (e.g., the implementation of a mandate of horizontal interoperability as codified in the EU DMA) focus on reducing the competitive advantages which large dominant platforms enjoy.[2] These measures promote contestability, which can be understood as "the ability of non-dominant firms to overcome barriers to entry and expand to the benefit of users" (Crémer et al., 2023), and are thought to benefit users by strengthening the competitive pressure which dominant platforms are exposed to.

Our analysis demonstrates that regulation which reduces a dominant platform's competitive advantage may have adverse effects on users. Simply put, this is because user migration away from a dominant platform incentivizes platforms to differentiate their content more, which adversely affects welfare—in particular, the dominant platform will then display more harmful content. Moreover, we show that there are profound complementarities between regulation that reduces a dominant platform's competitive advantage and initiatives that promote awareness regarding the adverse effects of harmful content. For example, reductions of a dominant platform's competitive advantage cannot raise user welfare if the share of users who disregard the effects of harmful content is large. This is because a dominant platform never has incentives to display non-harmful content in such settings. However,

[1] The evidence suggests that up to 60% of social media users experience such self-control problems.

[2] A mandate of horizontal interoperability as defined in the DMA specifies that users of a given platform must be able to interact with users of another competing platform. This reduces a dominant platform's competitive advantage by eliminating network effects.

initiatives that promote awareness regarding harmful content also feature trade-offs: While they can benefit all users, they may also increase a dominant platform's market share.

Formally, we consider a theoretical model in which there are two social media platforms that compete to attract users. The platforms simultaneously choose the share of harmful (yet engaging) content they show to users who join their platform. Based on the platforms' chosen harmful content shares, users decide which platform to join. One platform, which we refer to as the incumbent, has a competitive advantage over its rival, which we refer to as the entrant.[3] This means that, if both platforms display the same share of harmful content, any user who joins the incumbent obtains a higher utility. Importantly, this competitive advantage can be arbitrarily small.

Any platform has incentives to display a large share of harmful content because this induces the users who join this platform to spend more time on it, which increases the platform's revenues.[4] However, being exposed to harmful content decreases a users' utility. We refer to the time a user spends on a platform as the user's engagement level. Users are heterogeneous in the degree to which they internalize the adverse effects of being exposed to harmful content. A fixed share of all users is rational, while the rest are naive: Rational users take the utility costs of being exposed to harmful content into account when choosing which platform to join, while naive users do not.

The simultaneous presence of naive and rational users gives rise to the following key trade-off: Any platform wishes to maximize the engagement of users that join the platform, which incentivizes it to increase the share of harmful content it displays. However, raising the share of harmful content too much will cause rational users to leave the platform, which causes a discontinuous decrease in the platform's profits. In a monopoly setting, a platform will thus optimally display a small (respectively, large) share of harmful content if the share of rational users is large (respectively, small).

Different parameter combinations give rise to structurally different equilibria. There potentially exists an equilibrium in which all naive users join the incumbent and all rational users join the entrant (and vice versa), an equilibrium in which all users join the incumbent, and equilibria in which platforms play mixed strategies and users split between the platforms.

Our first main result is that user welfare is strictly larger in the equilibrium in which all users join the incumbent than in any other equilibrium. Intuitively, this holds by the

[3]We view the incumbent as the dominant platform in a given social media ecosystem. Such platforms enjoy a superior competitive position due to a larger user base (which benefits users of the platform due to network effects) or through superior access to user data.

[4]The business model of social media platforms relies on advertising. For example, in the last quarter of 2024, 96.69% of Meta's revenue came from advertising. Higher user engagement leads to more ad interactions, prompting platforms to design their algorithms to enhance engagement (Kamath et al., 2014).

following logic: In any equilibrium in which naive users join one platform and rational users join another platform, the platform which naive users join will optimally choose to display a maximal share of harmful content. This implies that naive users will obtain minimal utility, and that it is unviable for rational users to join the platform which naive users join. In turn, this implies that the platform which rational users join optimally offers them zero utility, since these users lack an outside option that would grant them positive utility. By contrast, all users must obtain strictly positive utility in the equilibrium in which all users join the incumbent. Otherwise, the entrant would deviate from the equilibrium by displaying a minimal level of harmful content, which induces rational users to join it.[5]

This insight establishes that reductions of a dominant platform's competitive advantage may have non-monotonic effects on user welfare: In the equilibrium in which all users join the incumbent (which emerges if the incumbent has a strong competitive advantage and the share of rational users is large), reductions of the incumbent's competitive advantage can increase user welfare. Intuitively, this is because such measures improve the entrant's ability to attract users, which means that it poses a stronger competitive threat. This incentivizes the incumbent to reduce the share of harmful content it displays to retain rational users, which benefits all users. However, if reductions of the incumbent's competitive advantage induce user migration away from the incumbent, this makes all users worse off in our framework.[6]

Our analysis also uncovers profound complementarities between regulation that reduces a dominant platform's competitive advantage and initiatives that promote awareness regarding the adverse effects of being exposed to harmful content. For example, improvements of the entrant's ability to generate utility for its users (which could be achieved, for example, by a mandate of horizontal interoperability) do not affect user welfare if the share of rational users is low. If the share of rational users is low, the incumbent displays maximal harmful content in equilibrium and naive users join this platform, while rational users join the entrant but always obtain zero utility. Thus, such policy measures do not raise user welfare, but only enable the entrant to retain rational users even when displaying more harmful content.

Relatedly, we demonstrate that the user-optimal outcome emerges if and only if the incumbent has no competitive advantage and the share of rational users is large enough. If the share of rational users is large enough, competition between symmetric platforms revolves

[5]This insight carries over when there is one incumbent platform with a competitive advantage and several other non-dominant platforms. If all users visit the incumbent but rational users obtain zero utility in equilibrium, one non-dominant platform would deviate from the equilibrium by attracting rational users.

[6]This insight extends even if platforms can extract utility through ad loads or through prices. If all users visit the incumbent, they must obtain positive utility, because they could be poached by another platform otherwise. In an equilibrium in which naive users visit one platform and rational users visit another platform, the platform which rational users visit can fully extract the surplus from these users.

around rational users, who always join the platform which displays a lower share of harmful content. In the unique equilibrium, both firms thus choose to display zero harmful content because they would not be joined by rational users otherwise (in equilibrium).

These results point to the importance of initiatives that raise awareness regarding the adverse effects of harmful content. However, such initiatives also feature trade-offs: While increases in the share of rational users can raise the utility of all users, these changes can also lead to increases in the market share of the incumbent by the following logic: If the share of rational users is small, the incumbent optimally displays a maximal amount of harmful content and foregoes rational users entirely. However, when the share of rational users becomes sufficiently large, the incumbent reduces the share of harmful content it shows to also attract rational users, which means that it will capture the entire market in equilibrium. Such levels of market dominance may have undesirable effects, particularly within dynamic contexts. Interestingly, regulation that enforces content moderation faces similar trade-offs. This is because such regulation makes it more profitable for the incumbent to attract all users by displaying a small share of harmful content.

Throughout our baseline analysis, we impose several simplifying assumptions to streamline the analysis. After presenting our main results, we show that our insights extend in the presence of network effects, when users can multi-home, and if rational and naive users on the same platform choose different levels of engagement. Moreover, our insights also apply in settings where there are no naive users but users have heterogeneous switching costs.

Finally, we note that the model's technology parameters and the share of users who are rational in our sense could be empirically estimated. This would enable sharper predictions regarding the likely equilibrium outcomes under different competitive scenarios.

**Related Literature:** Our paper contributes to the literature on platform competition and work on the optimal regulation of digital markets. To the best of our knowledge, we are the first to consider a model of platform competition in which asymmetric platforms choose how much harmful content to display to their users.

There is an extensive literature on platform competition (Rochet and Tirole, 2003; Armstrong, 2006; Anderson and De Palma, 2012; Bordalo et al., 2016; Prat and Valletti, 2022; Teh et al., 2023; Anderson and Peitz, 2023; Ekmekci et al., 2025). In contrast to most previous papers, we consider platforms that compete via content provision rather than pricing or ad loads. Our focus on content provision rather than ad loads builds on the insights of Brynjolfsson et al. (2024), who empirically document that exposure to advertisements has minimal effects on user welfare.

Our preference framework builds on the work of Ichihashi and Kim (2023) and Beknazar-Yuzbashev et al. (2024), who study the incentives of symmetric platforms to display harmful (yet engaging) content. Bhargava (2023) studies how competition by a platform which exogenously displays no harmful content affects an incumbent platform's incentives to display addictive content. Wickelgren and Gilo (2024) show how the provision of addictive content by a platform can deter entry. These papers follow recent contributions which suggest that digital platforms have incentives to provide content which increases users' engagement at the cost of their welfare (Rosenquist et al., 2021; Scott Morton and Dinielli, 2022).

Our model features boundedly rational users, which are largely absent from previous work on platforms (as discussed, for example, by Jullien et al. (2021)). An exception is Acemoglu et al. (2024), who study a model of digital advertising in which some users incorrectly interpret the information that ads convey about the quality of products. Importantly, Acemoglu et al. (2024) consider a different form of bounded rationality than we do and platforms in Acemoglu et al. (2024) do not choose how much harmful content to display.

A growing empirical literature documents the harmful effects of social media (Mosquera et al., 2020; Allcott et al., 2020; Horwitz et al., 2021; Braghieri et al., 2022).[7] A majority of social media users do not seem to take these adverse effects into account when deciding which platforms to join and how much time to spend consuming content: For example, Hoong (2021) and Allcott et al. (2022) document that a significant share of social media users suffer from such self-control problems.[8] Bursztyn et al. (2023) formalize that social media may adversely affect social welfare if users experience disutility from not participating.

A handful of papers study the effects of regulation regarding contestability. Kades and Scott Morton (2020) argue that interoperability is essential for healthy platform competition. Bourreau and Krämer (2023) and Dhakar and Yan (2024) show that a mandate of horizontal interoperability may reduce the share of users who multihome, which weakens competition, and may incentivize platforms to reduce the quality of their service.

**Outlook:** The rest of the paper proceeds as follows: In Section 2, we set up our theoretical model. Section 3 analyzes the monopoly benchmark. In Section 4, we characterize the competitive equilibria that emerge in our model. Section 5 studies a benchmark with perfect competition, and Section 6 discusses the policy implications of our results. We consider some extensions in Section 7 and conclude thereafter.

[7] The potential harmful effects of advertising have already been discussed by Becker and Murphy (1993).

[8] Algorithmic personalization further increases engagement at the potential cost of welfare (Guess et al., 2023; Beknazar-Yuzbashev et al., 2025).

# 2 Model

In this section, we lay out our model of platform competition.

*Players:* There is a unit mass of users indexed $i$, and two platforms $p \in \{E, I\}$ that users can join. We refer to the two platforms as the incumbent (in the sense of a market-dominating platform, indexed $I$) and the entrant (indexed $E$). Every user can only join one platform (in Section 7.1, we show that our results extend when users can multi-home).

*User choices:* Any user must decide which platform to join. A given user's platform choice is represented by a variable $j_i \in \{I, E, \emptyset\}$, where $j_i = I$ (respectively, $j_i = E$) specifies that the user joins the incumbent (respectively, the entrant). The choice of a user who does not join any platform is represented by $j_i = \emptyset$. We refer to the time that a user spends on a platform as the user's engagement level and denote this by the variable $e_i \in \mathbb{R}$.

*Platform choices:* Every platform chooses the share of harmful content it presents to every user who joins it, which we label $h_p \in [0, 1]$.

*User preferences and heterogeneity:* When a given user $i$ joins a platform $p$ and her engagement level is $e_i \in \mathbb{R}$, the utility she attains is given by $U_p(h_p, e_i)$. A user that does not join any platform obtains zero utility.[9]

Users differ in the extent to which they internalize the utility costs associated with exposure to harmful content. A fraction $\rho$ of users are rational, while the remaining share $1 - \rho$ are naive. We refer to this dimension of heterogeneity as the user's type $t_i \in \{n, r\}$. A user's type—rational or naive—is private information.

For expositional simplicity, we assume that rational users' and naive users' chosen engagement level on a given platform $p$ will be the same. Specifically, the engagement level of a user who joins a platform $p$ that displays a harmful content share $h_p$ is represented by a function $e_p^*(h_p)$. In Section 7.2, we consider an extension in which we allow for differences between the engagement levels of rational and naive users on a given platform.

A rational user maximizes her utility through the choice of $j_i$. Specifically, a rational user joins platform $l$ instead of platform $k$ only if $U_l^r(h_l) \geq U_k^r(h_k)$, where $U_p^r(h_p) := U_p(h_p, e_p^*(h_p))$.

Naive users join the platform on which they obtain higher perceived utility. We refer to naive users' perceived utility of joining platform $p$ as $U_p^n(h_p)$. Because naive and rational users' engagement levels on a given platform are the same, the true utility a naive user attains when joining platform $p$ is $U_p^r(h_p)$.

If platforms play pure strategies, user welfare is given by $\int_0^1 \left( \sum_{k \in \{I,E\}} \mathbb{1}[j_i = k] U_k^r(h_k) \right) \mathrm{d}i$.

[9] Our insights naturally extend to settings with negative outside options as in Bursztyn et al. (2023).

We further impose the following assumptions that are in line with our understanding of harmful content and user naivety in real-world social media platform markets:

**Assumption 1.** *The following assumptions hold:*

1. *The incumbent platform has a competitive advantage:* $U_I^r(h) > U_E^r(h)$ *and* $U_I^n(h) > U_E^n(h)$ *holds for all* $h \in [0,1]$.

2. *Exposure to harmful content decreases true utility: For both* $p \in \{I, E\}$*, the function* $U_p^r(h)$ *is strictly decreasing in* $h$*. Further,* $U_p^r(1) < 0 < U_p^r(0)$ *holds.*

3. *Naive users are drawn to harmful content: For both* $p \in \{I, E\}$*, the function* $U_p^n(h)$ *is strictly increasing in* $h$*.*

4. *Harmful content is more engaging: For both* $p \in \{I, E\}$*, the function* $e_p^*(h)$ *is strictly increasing in* $h$*.*

When discussing our microfoundation towards the end of this section, we provide parametric conditions under which these assumptions are satisfied. We abstract from the presence of network effects in our main analysis, and consider these in Section 7.3. Further, we consider settings in which users have heterogeneous switching costs in Section 7.4.

*Platform preferences:* Platform revenues are proportional to the engagement of users who join. Specifically, the revenue of platform $p \in \{E, I\}$ is given by the function

$$\Pi_p = \int_0^1 \mathbb{1}[j_i = p]\big(\mathbb{1}[t_i = r]\pi_p^r(e_p^*(h_p)) + \mathbb{1}[t_i = n]\pi_p^n(e_p^*(h_p))\big)di. \tag{1}$$

For every $p \in \{I, E\}$ and every $t \in \{r, n\}$, $\pi_p^t(x)$ is an increasing function. A platform may thus obtain different revenues from a naive user than from a rational user (fixing engagement).

*Timing:* The timing of the game is as follows: First, the two platforms simultaneously choose $h_E$ and $h_I$. After observing these choices, users decide which platform to join. Thereafter, utilities and profits are realized.

*Equilibrium:* Our equilibrium concept is a version of subgame-perfect equilibrium that accounts for naive users' behavior. A combination of users' and platforms' strategies is an equilibrium if and only if the following conditions jointly hold:

- For any $(h_E, h_I)$, a rational user's strategy maximizes her utility.
- For any $(h_E, h_I)$, a naive user's strategy maximizes her perceived utility.
- Each platform maximizes its revenue, given the other platform's and users' strategies.

**Microfoundation:** We now present an example of a simple preference framework, together with a particular specification of bounded rationality, which fits the aforementioned setting.

Any user who joins a platform $p$ and chooses engagement $e_i$ obtains the following utility:

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

where the parameters $\eta_p$ and $\theta_p$ characterize the sophistication of the platform's technology, i.e., to what extent the consumption of content raises a user's utility. For each unit of time spent on the platform, the user derives utility from consuming the platform's mix of harmful and non-harmful content (the first term). Further, she receives an engagement-independent benefit from joining the platform that increases with the share of non-harmful content this platform displays (the second term) and incurs a disutility from the harmful content she consumes (the third term).[10] The user also bears an opportunity cost of time spent on the platform (the fourth term). The utility-maximizing engagement level is $e_p^*(h_p) = \frac{(\eta_p - \theta_p)h_p + \theta_p}{2\gamma}$.

Given her optimal engagement, the utility a user obtains when joining a platform $p$ is:

$$U_p^r(h_p) = \frac{(h_p\eta_p + (1-h_p)\theta_p)^2}{4\gamma} + (1-h_p) - \delta h_p. \tag{3}$$

Rational users maximize their utility, i.e., join platform $p$ instead of $j$ if and only if $U_p^r(h_p) \geq U_j^r(h_j)$. Naive users disregard the adverse effects of being exposed to harmful content when choosing which platform to join: Specifically, they neglect the utility component $-\delta h_p$. Intuitively, $\delta > 0$ can be thought to capture the (long-term) costs of being exposed to harmful content, which naive users disregard. Naive users' perceived utility of joining platform $p$ is:

$$U_p^n(h_p) = \frac{(h_p\eta_p + (1-h_p)\theta_p)^2}{4\gamma} + (1-h_p) \tag{4}$$

This example fits our general specification under the following assumptions: Firstly, assuming that $\eta_I > \eta_E$ and $\theta_I > \theta_E$ implies that the incumbent has a competitive advantage. Secondly, setting $\delta$ appropriately large guarantees that $U_p^r(h_p)$ is decreasing in $h_p$ while $U_p^n(h_p)$ is increasing in $h_p$. Thirdly, $e_p^*(h_p)$ is an increasing function if $\eta_p > \theta_p$.

We emphasize that this is just an example of the settings which can be captured by our model. All results in Sections 4.1 and 4.2 are valid for our general setting. In Section 4.3, we build further intuition by re-considering the specific example we just laid out.

[10] The specification that users observe the platform's chosen harmful content share before deciding which platform to join can be interpreted as the build-up of reputation in a dynamic setting.

**Interpretation of our model:** We highlight key features of our framework and explain how it captures salient aspects of real-world social media platform markets.

*Model of social media platforms:* We study a model of social media platforms such as Instagram, TikTok, 𝕏, and BlueSky, consistent with the definition in Aridor et al. (2024). Users consume content curated by algorithmic feeds. Platform revenues are proportional to the total engagement generated by their users. Our model thus allows us to capture real-world scenarios of platform competition such as Instagram vs TikTok or 𝕏 vs BlueSky. We note that the model also applies to content creation platforms such as YouTube.

*Harmful content:* We understand harmful content as content that reduces users' utility but boosts engagement (points 2 and 4 of Assumption 1). This aligns with the definition of addictive content in Ichihashi and Kim (2023) and harmful yet engaging content in Beknazar-Yuzbashev et al. (2024). Intuitively, harmful content can be viewed as content which users would like to avoid ex ante, but are nevertheless drawn to.

*User heterogeneity:* Naive users are drawn to harmful content (Assumption 1, point 3), e.g., due to behavioural biases or addiction. The existence of such users is empirically documented by Braghieri et al. (2022). For the equilibrium analysis, it is immaterial whether naivete stems from genuine unawareness regarding the adverse effects of harmful content or from an inability to act in accordance with this awareness (e.g., due to self-control problems). However, this distinction becomes important when designing policy interventions that aim to raise the share of rational users as we define them. We return to this issue in Section 6.

*Technology:* The technology available to a platform determines its ability to provide users utility. This is based on the quality of the platform's algorithm, the size of its user base, and the platform's access to user data. Our assumption that the incumbent has a competitive advantage (Assumption 1, point 1) reflects the fact that a dominant position in a social media ecosystem grants a platform superior access to user data and a larger user base. Changes in the technology available to the platforms could stem from legislation such as a mandate for horizontal interoperability or from technological progress such as the expansion of AI.

*Advertising:* Even though we do not explicitly model advertising, the flexible specification of platform revenue we utilize can capture the role of advertising in reduced form.

*Content personalization:* Our insights naturally extend to settings in which platforms conduct third-degree personalization of the share of harmful content: Then, platforms play the game we laid out for each group of users with a given set of observable features on which personalization is based. Moreover, a platform that shows the same proportion of harmful content to all users may still display different content to each user, as the types of content that are harmful can differ across individuals. We elaborate more on this in Section 7.5.

# 3 Monopoly Solution

In this section, we briefly characterize the optimal behavior of a monopolist platform. This helps illustrate a fundamental trade-off that platforms also face under competition: Each platform seeks to maximize the engagement of users who join, because higher engagement grants the platform higher revenues. This incentivizes the platform to increase the share of harmful content it displays because this boosts the engagement level of its users, ceteris paribus. However, if this share becomes too large, rational users will opt not to join the platform. In that case, the platform foregoes any revenue from rational users.

Formally, suppose the incumbent platform is a monopolist. The monopolist's optimal approach takes one of two forms: It will either set an intermediate level of harmful content to ensure it is joined by all users, or set the maximal level of harmful content to maximize engagement by naive users. Formally, the monopolist will either optimally choose the harmful content share $h_I = \tilde{h}_I$, where $\tilde{h}_I$ is the level of harmful content at which a rational user is exactly indifferent between joining the platform or not, or choose the harmful content share $h_I = 1$, which maximizes engagement by naive users. Choosing the content level $h_I = 1$ is optimal for the monopolist if the share of rational users is small enough, i.e., if:

$$(1-\rho)\pi_I^n(e_I^*(1)) \geq \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)), \tag{5}$$

where $\tilde{h}_I \in [0,1]$ is the unique solution to the equation

$$U_I^r(\tilde{h}_I) = 0. \tag{6}$$

# 4 Equilibrium Analysis

## 4.1 Pure-strategy equilibria

In this subsection, we provide an analytical characterization of all equilibria in which both platforms play a pure strategy. For convenience, we refer to such equilibria as pure-strategy equilibria. We refer to the harmful content share chosen by the incumbent (respectively, the entrant) in equilibrium as $h_I^*$ (respectively, $h_E^*$).

Before formally defining the set of equilibrium candidates, it is useful to re-iterate the optimal behavior of users. Rational users join a given platform if and only if they obtain strictly positive utility when joining the platform (i.e., their participation constraint is satisfied) and if they prefer to join this platform instead of the platform's rival. Naive users join

a given platform if and only if their perceived utility from joining this platform is strictly positive (i.e., their participation constraint is satisfied), and if they prefer to join this platform instead of the platform's rival. We note that the true utility which rational and naive users obtain on a given platform is identical (if the platform plays a pure strategy), even though the naive users' perceived utility is different.

It is useful to characterize what would be optimal in terms of user welfare. The following Lemma characterizes the user-optimal outcome:

**Lemma 1** (User-optimal outcome)**.**
*User welfare is maximal if $h_I = 0$ and all users join the incumbent.*

Technically, this result follows from the fact that $U_I^r(h)$ is a decreasing function and the fact that the incumbent has a competitive advantage (see Assumption 1). Intuitively, the result holds because the true utility a user obtains on a given platform is maximal if this platform shows no harmful content, and since any user would obtain higher utility by joining the incumbent if both platforms display the same share of harmful content.

We continue the analysis by comparing equilibria in which all users join the incumbent to equilibria in which they do not.

**Proposition 1** (Harmful effects of endogenous differentiation)**.**
*In any pure-strategy equilibrium in which all users join the incumbent, the utility of all users is strictly larger than in any other pure-strategy equilibrium.*

The logic underlying this result is as follows: In any pure-strategy equilibrium in which the incumbent is joined by all users, rational users must obtain strictly positive utility by joining this platform. Otherwise, the entrant would deviate from the equilibrium by choosing a harmful content share at which rational users would obtain strictly positive utility by joining the entrant (which induces rational users to join the entrant, thereby granting it higher profits than it obtains in equilibrium).

By contrast, all rational users obtain zero utility in any equilibrium in which some users join the entrant and other users join the incumbent. To see this, suppose that all naive users join platform $l \in \{I, E\}$ and all rational users join platform $p \neq l$.[11] Because platform $l$ is only joined by naive users, it is optimal for this platform to set $h_l = 1$. This maximizes the engagement and the perceived utility of its users. Since $h_l = 1$, rational users would attain negative utility by joining platform $l$. In turn, this implies that rational users must, in equilibrium, obtain zero utility when joining platform $p$. Otherwise, the platform $p$ would

[11] In this example, all users play a pure strategy. In the proof, we demonstrate that the stated result holds true even if some users mix.

deviate by slightly increasing the share of harmful content it shows. This is because this deviation would leave the demand which this platform receives unaffected, but would increase the engagement of its users and hence, the platform's profits. Thus, rational users obtain zero utility in any pure-strategy equilibrium in which the entrant is joined by some users, i.e., strictly less utility than in an equilibrium in which all users join the incumbent.

Recall that rational and naive users who visit the same platform obtain the same utility. This implies that naive users also obtain strictly higher utility in any equilibrium in which all users visit the incumbent: In such an equilibrium, their utility is strictly positive, while their utility is negative in any other pure-strategy equilibrium because they are exposed to maximal harmful content.

This result provides a cautionary perspective on regulation that promotes user migration from a dominant platform to an entrant. This includes any regulation which reduces a dominant platform's competitive advantage, such as a mandate of horizontal interoperability.

To further characterize the possible pure-strategy equilibria that can emerge, we define some terminology regarding the levels of harmful content at which rational users would obtain exactly zero utility when joining a given platform. For either $p$, we define $\tilde{h}_p \in [0,1]$ as the unique solution to the following equation:

$$U_p^r(\tilde{h}_p) = 0. \tag{7}$$

Moreover, we define an object $\check{h}_I$ such that rational users are exactly indifferent between joining the incumbent and the entrant if the entrant chooses the harmful content share zero and the incumbent chooses the harmful content share $\check{h}_I$.[12] Thus, this object solves:

$$U_E^r(0) = U_I^r(\check{h}_I) \tag{8}$$

Having defined this, we are ready to characterize all pure-strategy equilibrium candidates:

**Proposition 2** (Pure-strategy equilibrium candidates)**.**
*There are three candidates for a pure-strategy equilibrium, namely:*

- *An equilibrium in which* $h_E^* = \tilde{h}_E$ *and* $h_I^* = 1$.
- *An equilibrium in which* $h_E^* = 1$ *and* $h_I^* = \tilde{h}_I$.
- *An equilibrium in which* $h_E^* = 0$, $h_I^* = \check{h}_I$, *and all users join the incumbent.*

[12]There exists a unique $\check{h}_I$ by the following arguments: The function $f(h) = U_E^r(0) - U_I^r(h)$ is strictly increasing, strictly negative at $h = 0$, and strictly positive at $h = 1$ (these facts hold by Assumption 1).

In the following, we refer to the first equilibrium candidate as the *naivety-focused equilibrium* and to the third equilibrium candidate as the *market dominance equilibrium*.

The logic underlying this equilibrium characterization result is as follows: There are four possible candidates for an equilibrium in which platforms play pure strategies: (i) Equilibria in which naive users join the incumbent and rational users join the entrant, (ii) equilibria in which naive users join the entrant and rational users join the incumbent, (iii) equilibria in which some users mix, and (iv) equilibria in which all users join the incumbent.

There is a unique candidate for an equilibrium in which naive users join the incumbent and rational users join the entrant. This result follows from previous arguments—if all naive users join the incumbent, this platform finds it optimal to set $h_I^* = 1$, which induces the entrant to set $h_E^* = \tilde{h}_E$ and extract all surplus from rational users. Similar arguments establish that there is a unique candidate for an equilibrium in which all naive users join the entrant and all rational users join the incumbent, and that $h_E^* = 1$ and $h_I^* = \tilde{h}_I$ must hold in this equilibrium. In any equilibrium in which some users mix, the incumbent must set the harmful content share $h_I^* = \tilde{h}_I$ and the entrant sets $h_E^* = 1$.

There is a unique candidate for an equilibrium in which all users join the incumbent. In such an equilibrium, rational users must be indifferent between joining the incumbent and the entrant.[13] In equilibrium, the entrant must set $h_E^* = 0$ to provide the strongest possible incentives for rational users to join it.[14] This implies that $h_I^* = \check{h}_I$ must hold, given that rational users must be indifferent between joining either platform in equilibrium.

In the following, we establish conditions under which the different equilibria exist.

**Proposition 3** (Pure-strategy equilibria: existence)**.**
*The existence regions for the pure-strategy equilibrium candidates are as follows:*

- *An equilibrium in which* $h_E^* = \tilde{h}_E$ *and* $h_I^* = 1$ *exists if and only if* $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \leq (1-\rho)\pi_I^n(e_I^*(1))$.
- *An equilibrium in which* $h_E^* = 1$ *and* $h_I^* = \tilde{h}_I$ *exists if and only if the conditions (i)* $U_E^n(1) \geq U_I^n(\tilde{h}_I)$ *and (ii)* $\frac{\rho}{1-\rho} \in \left[\frac{\pi_I^n(e_I^*(1))}{\pi_I^r(e_I^*(\tilde{h}_I))}, \frac{\pi_E^n(e_E^*(1))}{\pi_E^r(e_E^*(\tilde{h}_E))}\right]$ *jointly hold.*

[13] Suppose, for a contradiction, that there exists an equilibrium in which all users join the incumbent and rational users strictly prefer to join the incumbent. Then, the incumbent would strictly prefer to marginally raise its harmful content share, since all users still join the incumbent after the deviation (naive users obtain higher utility by joining the incumbent after the deviation) but the incumbent obtains higher engagement.

[14] To see this, note that $U_I^r(h_I^*) = U_E^r(0)$ must hold. If $U_I^r(h_I^*) > U_E^r(0)$, the incumbent would strictly prefer to slightly increase $h_I$. If $U_I^r(h_I^*) < U_E^r(0)$, the equilibrium cannot exist because the entrant would prefer to deviate by setting $h_E = 0$ and attracting all rational users. By implication, $h_E^* = 0$ must hold since rational users must be indifferent between joining either platform in equilibrium.

- *An equilibrium in which* $h_E^* = 0$ *and* $h_I^* = \check{h}_I$ *exists if and only if the conditions (i)* $U_E^n(1) \leq U_I^n(\check{h}_I)$ *and (ii)* $(1-\rho)\big(\pi_I^n(e_I^*(1)) - \pi_I^n(e_I^*(\check{h}_I))\big) \leq \rho\pi_I^r(e_I^*(\check{h}_I))$ *jointly hold.*

We can utilize this general result to build further intuition regarding the parameter regions for which the different equilibria emerge. For example, the naivety-focused equilibrium exists if and only if the share of rational users is small enough:

**Corollary 1** (The importance of awareness)**.**
*There exists a* $\underline{\rho} > 0$ *such that, if* $\rho < \underline{\rho}$*, there exists a unique equilibrium in which* $h_I^* = 1$ *and* $h_E^* = \tilde{h}_E$*.*

The intuition underlying this result is as follows: If $h_I^* = 1$ and $h_E^* = \tilde{h}_E$, then rational users optimally join the entrant and naive users join the incumbent. If the share of naive users is large enough, the incumbent finds it optimal to entirely focus on naive users and set a harmful content share of one. Then, the best choice the entrant has is to set $h_E = \tilde{h}_E$ and attract rational users, given that it can never attract naive users because of the incumbent's competitive advantage. If $\rho$ is small enough, setting $h_I^* = 1$ and $h_E^* = \tilde{h}_E$ is thus uniquely optimal for the platforms.

The results of Corollary 1 hint at profound complementarities between regulation that promotes contestability and initiatives that promote awareness regarding the adverse effects of harmful content: If the share of rational users is small enough, reducing the incumbent's competitive advantage cannot benefit users. There exists a unique equilibrium in which naive users always join the incumbent and consume maximal harmful content, while rational users join the entrant and attain utility zero. Thus, improving the entrant's ability to generate utility for its users will not benefit users, but will only enable the entrant to retain rational users even when showing them higher levels of harmful content. Conversely, reducing the incumbent's ability to generate utility will only reduce the utility of naive users.

## 4.2 Mixed-strategy equilibria

In this subsection, we characterize all possible equilibria in which at least one platform plays a mixed strategy, and refer to such equilibria as mixed-strategy equilibria. We refer to the distribution of the harmful content shares a platform $p$ chooses as $\Gamma_p$, and to the corresponding cumulative distribution function as $F_p$.

We begin by characterizing possible mixed-strategy equilibria in which all users join a given platform.

**Lemma 2** (Mixed-strategy equilibria with market dominance)**.**
*There exists no mixed-strategy equilibrium in which all users join the entrant with probability* 1*. In any mixed-strategy equilibrium in which all users join the incumbent with probability* 1*, the incumbent sets the harmful content share* $\check{h}_I$ *with probability* 1*.*

The logic underlying this result is as follows: There exists no equilibrium in which all users join the entrant with probability 1 (and the incumbent thus obtains zero profits) because the incumbent would deviate by setting a harmful content share of 1, which ensures that it makes positive profits. In any equilibrium in which all users join the incumbent with probability 1, the incumbent must play some harmful content share $h_I^*$ with probability 1.[15] In such an equilibrium, $U_E^r(0) = U_I^r(h_I^*)$ must hold. If $U_E^r(0) > U_I^r(h_I^*)$, the entrant would deviate from the equilibrium by setting a harmful content share of zero. If $U_E^r(0) < U_I^r(h_I^*)$, rational users would always strictly prefer to join the incumbent, which means the incumbent would prefer to slightly raise the share of harmful content it shows. Given that the equation $U_E^r(0) = U_I^r(h_I^*)$ has a unique solution (namely $\check{h}_I$), $h_I^* = \check{h}_I$ must hold.

From this, it follows that the utility which users obtain in any mixed-strategy equilibrium in which all users join the incumbent with probability 1 is the same as the utility which users obtain in the market dominance equilibrium. This is because all users join the incumbent and consume the harmful content share $\check{h}_I$ with probability 1 in either equilibrium.

We now establish that user welfare is smaller in any mixed-strategy equilibrium in which the entrant is joined by some users than in any equilibrium in which all users join the incumbent:

**Proposition 4** (Mixed-strategy equilibria: User welfare)**.**
*User welfare would be strictly smaller in any mixed-strategy equilibrium in which some users join the entrant with positive probability than in an equilibrium in which all users join the incumbent with probability* 1*.*

The logic underlying this result is as follows: In any equilibrium in which all users join the incumbent with probability 1, the incumbent sets the harmful content share $\check{h}_I$ and user welfare is exactly equal to $U_I^r(\check{h}_I)$. In any mixed-strategy equilibrium, the incumbent will never set a harmful content level strictly below $\check{h}_I$, because setting the harmful content level $\check{h}_I$ is always superior.[16] Moreover, the entrant will (by construction) always set a harmful content level weakly above 0. If a user joins the entrant when the entrant sets $h_E > 0$ or

[15]The incumbent does not mix because it would obtain the same demand for all harmful content shares it sets by specification of the equilibrium under consideration, but the engagement of users who join the incumbent's platform is increasing in the harmful content share it sets.

[16]To see this, note that $U_I^r(h_I) > U_E^r(h_E)$ holds for any $h_E$ if $h_I < \check{h}_I$. This holds by the definition of $\check{h}_I$.

a user joins the incumbent when the incumbent sets $h_I > \check{h}_I$, the user will obtain a utility level strictly below $U_I^r(\check{h}_I)$. By construction, one of these events must happen with positive probability in a mixed-strategy equilibrium in which some users join the entrant.[17]

## 4.3 Equilibrium predictions

In this section, we combine all previous insights to derive concrete equilibrium predictions by considering a particular parametric example. Specifically, we suppose that the utility of any rational user and the perceived utility of any naive user that joins a given platform $p$ take the same functional forms as in the microfoundation presented in Section 2, namely:

$$U_p^r(h_p) = \frac{(h_p\eta_p + (1-h_p)\theta_p)^2}{4\gamma} + (1-h_p) - \delta h_p, \tag{9}$$

$$U_p^n(h_p) = \frac{(h_p\eta_p + (1-h_p)\theta_p)^2}{4\gamma} + (1-h_p) \tag{10}$$

For expositional simplicity, we set $\pi_p^t(x) = x$ for both $p \in \{E, I\}$ and both $t \in \{n, r\}$.

Throughout the following analysis, we focus on parameter combinations for which $U_E^n(1) < U_I^n(\check{h}_I)$, i.e., for which the incumbent enjoys a relatively large competitive advantage. Considering such parameter combinations is particularly interesting because an equilibrium in which all users join the incumbent can only emerge if $U_E^n(1) \leq U_I^n(\check{h}_I)$. We begin by providing a more detailed characterization of the mixed-strategy equilibrium that can emerge under this specification.[18]

**Proposition 5** (Mixed-strategy equilibrium: Characterization)**.**
*Suppose $U_E^n(1) < U_I^n(\check{h}_I)$. There exists a unique candidate for a mixed-strategy equilibrium. In this equilibrium,* $\text{supp}\Gamma_I = [\underline{h}_I, \tilde{h}_I] \cup 1$ *and* $\text{supp}\Gamma_E = [\underline{h}_E, \tilde{h}_E]$ *must hold, where the values $\underline{h}_I$ and $\underline{h}_E$ must jointly solve the following equations:*

$$(1-\rho)e_I^*(1) = e_I^*(\underline{h}_I) \quad ; \quad U_E^r(\underline{h}_E) = U_I^r(\underline{h}_I). \tag{11}$$

*The equilibrium exists if and only if* $(1-\rho)e_I^*(1) \in \big(e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\big)$.

[17] To see why note the following: By definition of a mixed-strategy equilibrium, the incumbent must either set a harmful content above $\check{h}_I$ with positive probability or the entrant must set a harmful content level above 0 with positive probability. Moreover, a platform must obtain positive demand at any harmful content share it sets in equilibrium, which implies that some users will join the incumbent (respectively, the entrant) when this platform sets a harmful content level strictly above $\check{h}_I$ (respectively, above 0).

[18] A detailed explanation regarding the derivation of this equilibrium can be found in Section B.2 of the online appendix.

For convenience, we now provide a visualization of the supports of $\Gamma_I$ and $\Gamma_E$ in the mixed-strategy equilibrium under consideration. A round circle at a harmful content share indicates that the distribution of harmful content shares has an atom at this point:

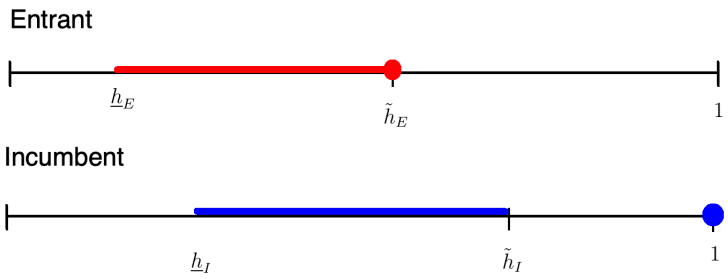


Figure 1: Visualization: Mixed-strategy equilibria

We now establish equilibrium existence and uniqueness:

**Proposition 6** (Equilibrium existence and uniqueness)**.**
*For any parameter combination at which $U_E^n(1) < U_I^n(\check{h}_I)$, there exists a unique equilibrium.*

This result follows from the insights of Propositions 3 and 5. If the share of rational users is small (i.e., $(1-\rho)e_I^*(1) \geq e_I^*(\tilde{h}_I)$), the unique equilibrium that exists is the pure-strategy equilibrium in which $h_I^* = 1$ and $h_E^* = \tilde{h}_E$. If the share of rational users is at an intermediate level, i.e., $(1-\rho)e_I^*(1) \in \left(e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\right)$ holds, the unique equilibrium that exists is the mixed-strategy equilibrium we characterized in the previous proposition. If the share of rational users is large (i.e., $(1-\rho)e_I^*(1) \leq e_I^*(\check{h}_I)$), the unique equilibrium that exists is the pure-strategy equilibrium in which $h_I^* = \check{h}_I$, $h_E^* = 0$, and all users join the incumbent.

We visualize these insights in the following figure. We consider a particular parametric examples in which $\theta_I = 3, \eta_I = 4$, $\gamma = 0.25$, and $\delta = 20$. Every graph corresponds to a given level of $\theta_E \in \{0.5, 1.5\}$. We plot different values of $\rho \in [0,1]$ on the x-axis of every graph and different values of $\eta_E \in [2, 2.5]$ on the y-axis of every graph. All parameter combinations we consider satisfy Assumption 1. Blue points indicate (for a given parameter combination) that the pure-strategy equilibrium $(h_I^*, h_E^*) = (1, \tilde{h}_E)$ is the unique equilibrium that exists. Yellow points indicate that the mixed-strategy equilibrium characterized in Proposition 5 is the unique equilibrium that exists. Green points indicate that the pure-strategy equilibrium in which all users join the incumbent is the unique equilibrium that exists.

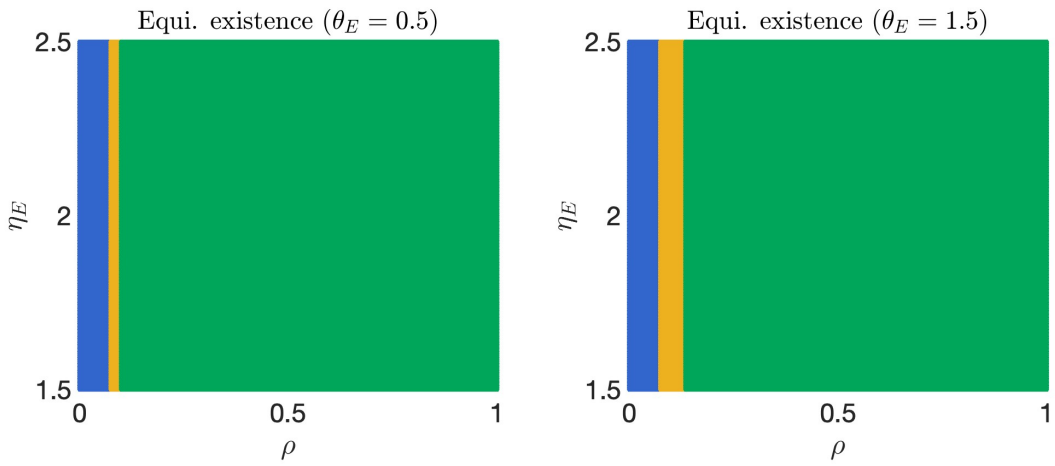


Figure 2: Equilibrium existence regions

The size of the parameter region for which the equilibrium in which $(h_I^*, h_E^*) = (1, \tilde{h}_E)$ exists is independent of the entrant's technology. However, the parameter region for which the equilibrium in which all users join the incumbent exists shrinks when $\theta_E$ increases. Intuitively, this is because increases of $\theta_E$ improve the entrant's ability to attract users, thereby making it less feasible for the incumbent to sustain an equilibrium in which all users join it. Formally, increases of $\theta_E$ imply that $U_E^r(0)$ increases, which means that $\check{h}_I$, i.e. the level of harmful content at which rational users would be indifferent between the entrant and the incumbent if the former sets $h_E = 0$ and the latter sets $h_I = \check{h}_I$, decreases. This reduces the profits which the incumbent obtains in the equilibrium in which all users join it, which increases the profitability of deviations from this equilibrium.

This analysis gives rise to three further interesting observations: Firstly, increases of $\theta_E$ may induce both platforms to set strictly higher harmful content shares and may reduce users' utility. To see this, consider a $\rho \approx 0.1$ and two $\theta_E^1, \theta_E^2$ such that $\theta_E^2$ lies just above $\theta_E^1$, the market dominance equilibrium in which $(h_I^*, h_E^*) = (\check{h}_I, 0)$ emerges if $\theta_E = \theta_E^1$, and the mixed-strategy equilibrium emerges if $\theta_E = \theta_E^2$. Since $\theta_E^1$ lies just below $\theta_E^2$, the level of $\check{h}_I$ is essentially the same for both parameter combinations. As the market transitions from the market dominance equilibrium to the mixed-strategy equilibrium in response to an increase of $\theta_E$, the incumbent now plays $\check{h}_I$ with probability strictly below 1. In addition, the entrant now plays harmful content levels strictly above 0 with probability 1. Thus, the expected harmful content shares of both platforms increases. This effect, together with the fact that some users now join the entrant, implies that the utility of all users falls.

Secondly, increases of $\rho$ go along with increases in the market share of the incumbent. This suggests that initiatives which attempt to improve user welfare in social media platform

markets by increasing the share of rational users (as we define them) may face a fundamental trade-off, because they induce the incumbent to obtain a more dominant position in equilibrium. This may have adverse effects by creating additional barriers to entry through network effects and by improving the incumbent's access to user data.

Thirdly, the previous analysis suggests that regulation on content moderation, which can be understood as the imposition of an upper bound on the share of harmful content platforms can display, may also increase the market share of the incumbent. This is because the region for which the market dominance equilibrium exists expands after the imposition of an exogenous upper limit on the harmful content share the incumbent platform can set. Intuitively, this is because deviations from this equilibrium (which would always be to the highest implementable harmful content share) become less profitable for the incumbent in response to this policy, which makes the equilibrium easier to sustain. Thus, regulation on content moderation may feature the same trade-offs as initiatives that promote awareness regarding the adverse effects of harmful content: Both increase the welfare of users, but may strengthen the dominant position of an incumbent platform.

# 5 Leveling the Technological Playing Field

In this section, we consider a benchmark in which the incumbent no longer has a competitive advantage and there are no network effects (which could be achieved, for example, by the implementation of a mandate to interoperability). Formally, we consider the parametric example from Section 4.3 and assume that $\eta_I = \eta_E = \eta$ and $\theta_I = \theta_E = \theta$ holds.

In such settings, the user-optimal outcome emerges if there are enough rational users. We formalize this in the following proposition.

**Proposition 7** (Leveling the playing field)**.**
*Suppose the incumbent has no competitive advantage. There is a* $\rho^* \in (0,1)$ *such that, if* $\rho > \rho^*$*, there exists a unique pure-strategy equilibrium in which* $h_I^* = h_E^* = 0$*.*

If $\rho$ is large enough, there exists an equilibrium in which all platforms display zero harmful content and all users join either platform with probability 0.5. If any platform deviates by increasing the share of harmful content, it will not be joined by rational users anymore. Thus, the most profitable deviation for a platform is to set the harmful content level $h = 1$, since this maximizes the perceived utility and the engagement of naive users. Such a deviation is not profitable if the share of rational users is large enough. Similar arguments establish that there exists no equilibrium in which any platform chooses a positive harmful content share.

It is worthwhile to clarify why this equilibrium does not exist when the incumbent has a competitive advantage. If both platforms set $h_E^* = h_I^* = 0$, then all users prefer to join the incumbent, since this delivers higher utility. But then, the entrant would deviate by setting a harmful content level of 1, given that this induces naive users to join it.

The results of this section can further guide the policy debate regarding the optimal regulation of social media platform markets: While reductions of the competitive advantages which dominant platforms enjoy may have adverse effects on users, eliminating such competitive advantages entirely will lead to the emergence of the user-optimal outcome if there is sufficient awareness regarding the adverse effects of harmful content.

# 6 Policy Implications

Social media platforms are under increasing regulatory scrutiny. In this section, we interpret our theoretical results in light of ongoing regulatory debates.

In platform markets, policymakers have traditionally advocated for measures that reduce a dominant platform's competitive advantage. Such measures are thought to unfold beneficial effects by strengthening the competitive pressure which dominant platforms are exposed to. One prominent example of such a policy proposal is the establishment of a mandate to horizontal interoperability, which improves small platforms' ability to generate utility for their users by eliminating network effects.

Our analysis reveals that such policy interventions may have non-monotonic effects in social media markets because of the presence of users which neglect the adverse effects of harmful content. Nevertheless, the user-optimal outcome can emerge if the competitive advantage of a dominant platform is eliminated entirely, but only if there is sufficient awareness regarding the adverse effects of being exposed to harmful content.

Recent regulatory approaches by the EU reflect some appreciation of the fact that this type of awareness is important. Notably, the Digital Services Act (DSA) requires platforms to provide transparency around the functioning of their recommendation algorithms.[19] There are two further (potentially complementary) ways of increasing the share of users who are rational by our definition. First, one could make users aware of the adverse effects of being exposed to harmful content. Second, one could provide users with tools to help manage and self-regulate their social media consumption. The former approach is sensible if a majority

[19]The DSA mandates that "online platforms should consistently ensure that recipients of their service are appropriately informed about how recommender systems impact the way information is displayed, and can influence how information is presented to them."

of users are truly unaware of the adverse effects of being exposed to harmful content, while the latter approach is effective if a large share users are already aware of these issues, but do not act on their knowledge. The empirical evidence suggests that both approaches may be effective: 34% of Americans report that social media is harmful to their mental health Statista (2025), while Bursztyn et al. (2023) estimate that over 50% of users experience a net utility loss because of their social media usage.

# 7 Extensions

## 7.1 Multi-homing

In this subsection, we show that our main insights continue to hold when users are allowed to multi-home. Formally, we consider the following model: At the beginning of the game, the platforms simultaneously choose their harmful content shares. After observing the harmful content shares chosen by the platforms, each user can choose to join the incumbent, the entrant, neither platform, or both platforms (i.e., to multi-home). This last option is new. A user who multi-homes and allocates engagement levels $e_E$ and $e_I$ to the entrant and the incumbent, respectively, obtains the following utility:

$$\sum_{j\in\{E,I\}} \big(\eta_j h_j + \theta_j(1-h_j)\big)e_j + 0.5(g_E - \delta h_E) + 0.5(g_I - \delta h_I) - \gamma\big(e_I + e_E\big)^2 \tag{12}$$

The perceived utility a naive user who multi-homes and allocates engagement levels $e_E$ and $e_I$ to the entrant and the incumbent, respectively, is:

$$\sum_{j\in\{E,I\}} \big(\eta_j h_j + \theta_j(1-h_j)\big)e_j + 0.5g_E + 0.5g_I - \gamma\big(e_I + e_E\big)^2 \tag{13}$$

The utility which users obtain when joining a platform $p$ is given by equation (3). The perceived utility which naive users obtain when joining a platform $p$ is given by equation (4). Rational and naive users choose their engagement levels to maximize their utility. Everything else is as in the baseline model.

We now characterize the equilibria that emerge under multi-homing.

**Proposition 8** (Multi-homing)**.**
*In any pure-strategy equilibrium in which some users multi-home and some users choose positive engagement levels on the entrant's platform, all users obtain weakly negative utility.*

This result holds by the following logic: In any equilibrium in which some users multi-home,

the users which multi-home must choose zero engagement on one platform: In a hypothetical equilibrium in which some users multi-home and spend time on both platforms, the entrant must set a larger harmful content share than the incumbent (else, users would only spend time on the incumbent platform). But this means that the user would strictly prefer to only join the incumbent, since she is indifferent between spending time on either platform and the incumbent provides more good content (and thus, the engagement-independent utility a user obtains would be larger if she only joins the incumbent). Hence, there exists no equilibrium in which users multi-home and spend time on both platforms.

This implies that, in any equilibrium with multi-homing (and in which some users choose positive engagement on the entrant platform), rational and naive users must obtain weakly negative utility. To see this, note firstly that there exists no equilibrium with multi-homing in which all users multi-home or only rational users multi-home. Moreover, the familiar logic from the baseline analysis applies in any equilibrium in which naive users multi-home: Suppose rational users join platform $p$. Then, platform $l$ will only obtain profits from multi-homing naive users in equilibrium, which implies that it optimally chooses the maximum harmful content share. But then, platform $p$ finds it optimal to extract all surplus from rational users. Taken together, these arguments imply the stated property.

These results indicate that our key insights go through even when users can multi-home: All users are strictly better off in an equilibrium in which all users join the incumbent than in any equilibrium with multi-homing in which users allocate positive engagement levels to the entrant. Moreover, all results pertaining to equilibria without multi-homing (from the baseline analysis) extend by construction.

## 7.2 Differences in engagement

In this subsection, we consider a model that is entirely analogous to the model we presented in Section 2, with one exception: We now allow the engagement levels of rational and naive users on a given platform to differ. Specifically, we denote the engagement choices of rational users on a platform $p$ by $e_p^r(h_p)$ and the engagement choices of naive users by $e_p^n(h_p)$. As before, we impose that parts 1 and 3 of Assumption 1 hold. To account for differences in engagement, we further impose that the function $e_p^t(h_p)$ is strictly increasing in $h$ for both types $t \in \{r, n\}$ and both platforms $p \in \{E, I\}$ and that the function $U_p(h, e_p^t(h))$ is strictly decreasing in $h$ for both $p \in \{I, E\}$ and both $t \in \{r, n\}$.

We show that the key prediction from the baseline analysis extends verbatim:

**Proposition 9** (Engagement differences)**.**
*In any pure-strategy equilibrium in which all users join the incumbent, the utility of all users is strictly larger than in any other pure-strategy equilibrium.*

The logic underlying this result is as in the baseline analysis: In any pure-strategy equilibrium in which all users join the incumbent, rational users must obtain strictly positive utility. In any pure-strategy equilibrium in which naive users join one platform and rational users join another platform, the platform which naive users join displays maximal harmful content. This means that rational users would obtain zero utility in such an equilibrium, i.e., obtain strictly lower utility than in a pure-strategy equilibrium in which all users join the incumbent.

It remains to argue why naive users obtain strictly larger utility in any pure-strategy equilibrium in which all users join the incumbent: The key reason is that they are exposed to a low amount of harmful content if all users join the incumbent, while they are exposed to maximal harmful content in any other pure-strategy equilibrium. Because exposure to harmful content decreases a user's utility and since the incumbent has a technological advantage, naive users are thus strictly worse off in any pure-strategy equilibrium in which some users join the entrant.

While our results regarding pure-strategy equilibria extend verbatim, we note that our welfare result pertaining to mixed strategy equilibria may not always carry over if there are engagement differences between rational and naive users. The result in Proposition 4, namely that the utility of all users is strictly larger in any equilibrium in which all users join the incumbent with probability 1, is based on the fact that rational users obtain strictly larger utility in any equilibrium in which all users join the incumbent with probability 1. This directly extends to naive users if there are no engagement differences between rational and naive users on a given platform, since rational and naive users who join the same platform would obtain the same utility. If there are engagement differences, this may not be true.

## 7.3 Network effects

In this subsection, we integrate the possibility of network effects into our baseline model. Specifically, we assume that the utility which rational users attain by joining platform $p$ is given by

$$U_p^r(h_p, s_p) = \frac{(h_p\eta_p(s_p) + (1-h_p)\theta_p(s_p))^2}{4\gamma} + (1-h_p) - \delta h_p, \tag{14}$$

where $h_p$ is the harmful content share chosen by this platform and $s_p$ is the share of users who join this platform in equilibrium. The functions $\eta_p(s_p)$ and $\theta_p(s_p)$ characterize the sophistication of the platform's technology.

The perceived utility which naive users attain when joining platform $p$ is given by

$$U_p^n(h_p, s_p) = \frac{(h_p\eta_p(s_p) + (1-h_p)\theta_p(s_p))^2}{4\gamma} + (1-h_p). \tag{15}$$

For simplicity, we assume that the engagement levels of naive users and rational users are given by the same function $e_p^*(h_p, s_p)$. Everything else is as in the baseline model. Moreover, we adopt the following assumption:

**Assumption 2.** *The following assumptions hold:*

1. *For any $h \in [0,1]$ and any $s \in [0,1]$, $U_I^r(h,s) > U_E^r(h,s)$ and $U_I^n(h,s) > U_E^n(h,s)$ hold.*
2. *For both $p \in \{I, E\}$ and any $s_p \in [0,1]$, the function $U_p^r(h, s_p)$ strictly decreases in $h$, $U_p^r(1, s_p) < 0$ holds, and $U_p^r(0, s_p) > 0$ holds.*
3. *For both $p \in \{I, E\}$, $\frac{\partial U_p^n(h,s_p)}{\partial h} > 0$ holds for all $h \in [0,1]$ and any $s_p \in [0,1]$.*
4. *For both $p \in \{I, E\}$ and any $s_p \in [0,1]$, the function $e_p^*(h, s_p)$ strictly increases in $h$.*

This assumption can be understood as an analogue of Assumption 1 that accounts for network effects. We refer to the model we have just described as the *network effects model*.

In the following, we show that our key equilibrium prediction from the baseline analysis extends. Importantly, we restrict attention to equilibria in which the network size of the incumbent is larger (i.e. $s_I > s_E$), which is a natural restriction under the characterization of the technology available to the entrant and the incumbent:

**Proposition 10** (Network effects)**.**
*Restrict attention to pure-strategy equilibria in which the network size of the incumbent is larger. In any equilibrium in which all users join the incumbent, the utility of all users is strictly larger than in any equilibrium in which some users join the entrant.*

Thus, the key prediction from the baseline analysis extends. The underlying logic is identical.

## 7.4 Captive users

In this subsection, we consider an extension with rational users and users that are captive to the incumbent (instead of rational users and naive users). This constitutes an important

robustness check, given that social media platform markets are characterized by significant barriers to migration, which makes (some) users effectively captive to an incumbent platform. Thus, users which do not seek to join a platform that provides less harmful content might simply be captive.

We show that the key equilibrium prediction we obtain in this extension is analogous to the prediction from the baseline model. Moreover, reducing the barriers to migration can (perhaps counterintuitively) raise the market share of the incumbent platform.

Formally, we consider a model that is entirely analogous to the model outlined in Section 2, with one exception: A share $1-\rho$ of all users is captive to incumbent. As before, a share $\rho$ of all users is rational as defined in the baseline model. The chosen engagement levels of rational and captive users are identical and captured by the function $e_p^*(h_p)$. A user who is captive to the incumbent joins the incumbent and devotes engagement level $e_I^*(h_I)$ if the incumbent sets the harmful content share $h_I$. For simplicity, we assume that $\pi(x) = x$. Everything else is as in the baseline model. We refer to the model we just laid out as the captive users model.

**Proposition 11** (Captive users)**.**
*Consider the captive users model:*

- *In any pure-strategy equilibrium in which all users join the incumbent, all users obtain strictly positive utility. In any pure-strategy equilibrium in which some users join the entrant, all users obtain weakly negative utility.*
- *There exist $\rho^1$ and $\rho^2$, with $\rho^1 < \rho^2$, such that the market share of the incumbent is $1-\rho$ if $\rho < \rho^1$ and is equal to 1 if $\rho > \rho^2$.*

## 7.5 Personalized content

Our insights naturally extend to settings in which any platform conducts third-degree personalization of the share of harmful content it displays. Consider a setting in which both platforms observe public information about each user and suppose that this information is informative about whether a user is naive or rational. Then, there are segmented markets—in particular, platforms play the game we laid out for each segment of users with a given set of observable features on which personalization is based. Within each segment, the equilibrium predictions of the model extend verbatim.

Therefore, the key insights of our main analysis carry over: Firstly, all users must be strictly better off in any equilibrium in which all users visit the incumbent than in any other equilibrium. To see this, consider an equilibrium in which some users visit the entrant. Then,

the utility which any such user attains is lower than the utility she would attain if all users in her segment visit the incumbent in equilibrium. Secondly, the effects of initiatives that promote awareness regarding harmful content are the same in every segment and analogous to the effects we discussed previously.

We also note that, even if firms cannot conduct third-degree personalization of the harmful content share, a platform may still display different content to different users. This is because what constitutes harmful content may vary across users. In addition, mixed-strategy equilibria naturally emerge in the settings we considered. In any such equilibrium, different users on a given platform will be shown different shares of harmful content.

# 8 Conclusion

Social media platforms are a substantial societal issue: They foster political polarization as well as the emergence of mental health issues. To a large extent, these adverse effects stem from the fact that any social media platform has incentives to display a large share of addictive content (which is harmful to its users) because this boosts the time that users spend on the platform. Importantly, many users of social media platforms do not seem to internalize the adverse effects of being exposed to harmful content when deciding which platforms to join. In this paper, we study how competition between social media platforms is shaped by the presence of users which are naive in this sense. We demonstrate that reductions of a dominant platform's competitive advantage, which is the standard regulatory approach for platform markets, has non-monotonic effects. Moreover, there are profound complementarities between regulation that reduces a dominant platform's competitive advantage and initiatives that promote awareness regarding the adverse effects of harmful content.

# A Proofs:

**Proof of Lemma 1:** In the user optimum, all users must join the incumbent because the incumbent has a competitive advantage. If all users join the incumbent, their utility is maximized if $h_I = 0$. ■

**Proof of Proposition 1:**

**Part 1:** In any pure-strategy equilibrium in which the entrant is joined by some users, rational users obtain zero utility and naive users obtain weakly negative utility.

Firstly, consider an equilibrium in which both platforms and all users play a pure strategy. Suppose all rational users join platform $p$ and all naive users join platform $l \neq p$.

In equilibrium, platform $l$ must set $h_l = 1$. Suppose, for a contradiction, that $h_l < 1$. In equilibrium, naive users must weakly prefer this platform, and rational users join the other platform. The participation constraint of naive users is always slack. If platform $l$ deviates by setting $h_l = 1$, this will raise the utility that naive users attain on platform $l$, so they would still choose to join this platform after the deviation. Moreover, rational users do not join platform $l$ in equilibrium. Thus, the deviation raises the total engagement that platform $l$ receives without reducing its demand. Hence, the deviation is profitable, a contradiction.

The fact that $h_l = 1$ must hold means that rational users would attain negative utility when joining platform $l$ (by Assumption 1). It also implies that naive users obtain negative utility in equilibrium (again, by Assumption 1).

In equilibrium, rational users must obtain zero utility. Suppose, for a contradiction, that rational users attain strictly positive utility by joining platform $p$. Then, platform $p$ would find it optimal to marginally increase the share of harmful content it displays. After the deviation, rational users would still strictly prefer to join platform $p$ (since rational users would obtain negative utility by joining platform $l$), but the platform obtains higher engagement from all rational users who join it. If naive users would also join the platform after the deviation, the deviation become even more profitable. Hence, the deviation is profitable, which is a contradiction.

Secondly, consider equilibria in which both platforms play a pure strategy and some users play a mixed strategy. Suppose naive users mix (which means they must be indifferent between both platforms). This means that rational users cannot mix.[20] Suppose rational

[20]Suppose naive users are indifferent. Because the incumbent has a competitive advantage, $h_I < h_E$ must

users join platform $p$. This means that platform $l$ is only joined by naive users, so it will optimally set $h_l^* = 1$. By implication, this implies that $h_p^* = \tilde{h}_p$ must hold. All users who join platform $p$ obtain zero utility in equilibrium, while all users who join platform $l$ obtain negative utility.

Finally, note that there exists no equilibrium in which platforms play a pure strategy and rational users mix. In such an equilibrium, naive users must strictly prefer to join some platform. Suppose naive users join platform $p$. Then, platform $l$ would strictly prefer to marginally reduce the harmful content share it offers, because all rational users join platform $l$ after the deviation.

**Part 2:** In any pure-strategy equilibrium in which all users join the incumbent, all users obtain strictly positive utility.

Note that the entrant can always guarantee that any rational user who joins it obtains positive utility by setting a harmful content share in a small open interval above zero. If the entrant sets such a harmful content share and is joined by rational users, it obtains strictly positive profits.

Suppose, for a contradiction, that rational users joins the incumbent but attain utility zero. Then, the entrant would deviate by setting a harmful content share in a small open interval above zero. After the deviation, rational users would join the entrant and choose positive engagement. Thus, the deviation is profitable because it enables the entrant to obtain positive profits (while it obtains zero profits in equilibrium). This is a contradiction.

Hence, rational users obtain strictly positive utility when joining the incumbent. Naive users who join the incumbent obtain the same utility. This completes the proof. ■

**Proof of Proposition 2:**

**Part 1:** Characterizing equilibria in which all rational users join a platform $p$ and all naive users join a platform $l \neq p$.

If all naive users join the incumbent, $h_I^* = 1$ must hold by previous arguments. Moreover, rational users must obtain zero utility in equilibrium, which implies that $h_E = \tilde{h}_E$ must hold.

If all naive users join the entrant, $h_E^* = 1$ must hold by previous arguments. Moreover, rational users must obtain zero utility in equilibrium, which implies that $h_I = \tilde{h}_I$ must hold.

These are the first two equilibrium candidates from the proposition.

hold. This means that rational users strictly prefer to join the incumbent.

**Part 2:** There is one candidate for an equilibrium in which all users join the incumbent. In such an equilibrium, $h_I^*$ and $h_E^*$ must jointly satisfy $h_E^* = 0$ and $U_E^r(0) = U_I^r(h_I^*)$.

Consider an equilibrium in which the incumbent is joined by all users. We say that a given user type's incentive constraint is satisfied if such users prefer to join the incumbent.

In equilibrium, the incentive constraint of rational users must bind. Suppose, for a contradiction, that it is slack. By previous arguments, rational users must obtain positive utility by joining the incumbent in equilibrium. But then, the incumbent would prefer to slightly raise the share of harmful content it displays (since all users will still strictly prefer to join the incumbent after the deviation), a contradiction.

This implies that $U_I^r(h_I^*) = U_E^r(0)$ must hold in equilibrium. To see this, note that $h_E = 0$ maximizes $U_E^r(h_E)$ by our assumptions. If $U_I^r(h_I^*) > U_E^r(0)$, the incentive constraint must be slack, a contradiction. If $U_I^r(h_I^*) < U_E^r(0)$, the entrant would prefer to deviate by setting $h_E = 0$, and all rational users would then join the entrant. These arguments establish that $h_I^* = \check{h}_I$ must hold. It follows that $h_E^* = 0$ must hold because the incentive constraint of rational users must bind.

**Part 3:** There exists no equilibrium in which all users join the entrant.

Suppose such an equilibrium exists. Then, the incumbent would deviate by setting $h_I = h_E^*$. After the deviation, all users strictly prefer to join the incumbent, which makes the deviation profitable.

**Part 4:** In any equilibrium in which some users mix, $h_E^* = 1$ and $h_I^* = \tilde{h}_I$ must hold.

By previous arguments, rational users cannot mix in equilibrium. Suppose naive users mix in equilibrium. Then, rational users cannot be indifferent. Suppose rational users join platform $p$. Then, platform $l \neq p$ must set $h_l^* = 1$ because it is just joined by naive users in equilibrium. Thus, $l = E$ must hold, because naive users cannot be indifferent otherwise. Further, $h_I^* = \tilde{h}_I$ must hold—else, the incumbent would strictly prefer to raise $h_I$. ■

**Proof of Proposition 3:**

**Equilibrium candidate 1:** An equilibrium in which $h_E^* = \tilde{h}_E$ and $h_I^* = 1$ exists if and only if $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \leq (1-\rho)\pi_I^n(e_I^*(1))$.

If platforms play these strategies, naive users prefer to join the incumbent. Rational users

strictly prefer to join the entrant because they obtain zero utility by joining the incumbent.

The entrant has no profitable deviations. By reducing $h_E$, it cannot attract more naive users, and will reduce engagement by rational users (which makes such deviations unprofitable). If it deviates by setting $h_E \in (\tilde{h}_E, 1]$, it will no longer be joined by rational users. Moreover, naive users will always strictly prefer to join the incumbent since $U_E^n(h_E)$ attains its maximum at $h_E = 1$ and $U_E^n(1) < U_I^n(h_I^*)$. Thus, all deviations $h_E \in (\tilde{h}_E, 1]$ are strictly unprofitable for the entrant.

Now consider possible deviations for the incumbent. All deviations $h_I \in (\tilde{h}_I, 1)$ cannot be profitable, since the incumbent would obtain the same demand as in equilibrium, but lower engagement. Within the interval $h_I \in [0, \tilde{h}_I]$, the most profitable deviation would be to $\tilde{h}_I$: When choosing any $h_I < \tilde{h}_I$, the incumbent would obtain weakly lower demand than when setting $h_I = \tilde{h}_I$ (since $\tilde{h}_I > \tilde{h}_E = h_E^*$ and all naive users would thus join the incumbent if it sets $\tilde{h}_I$), and lower engagement than when setting $\tilde{h}_I$. Thus, the most profitable deviation for the incumbent would be to set the harmful content share $\tilde{h}_I$. This deviation is not profitable under the stated condition. To see why this holds true, note that only naive users (with share $1 - \rho$) join the incumbent in equilibrium, while all users would strictly prefer to join the incumbent under the deviation (since $\tilde{h}_E < \tilde{h}_I$). Thus, the equilibrium profits are $(1-\rho)\pi_I^n(e_I^*(1))$, which must be above the deviation profits $\rho\pi_I^r(e_I^*(\tilde{h}_I))+(1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$.

**Equilibrium candidate 2:** An equilibrium in which $h_E^* = 1$ and $h_I^* = \tilde{h}_I$ exists if the conditions (i) $U_E^n(1) \geq U_I^n(\tilde{h}_I)$ and (ii) $\frac{\rho}{1-\rho} \in \left[\frac{\pi_I^n(e_I^*(1))}{\pi_I^r(e_I^*(\tilde{h}_I))}, \frac{\pi_E^n(e_E^*(1))}{\pi_E^r(e_E^*(\tilde{h}_E))}\right]$ jointly hold.

In equilibrium, rational users strictly prefer to join the incumbent. Naive users prefer to join the entrant if and only if condition (i) holds.

Firstly, consider possible deviations by the entrant. In equilibrium, it is joined by naive users and obtains profits of $(1-\rho)\pi_E^n(e_E^*(1))$. All deviations $h_E \in (\tilde{h}_E, 1)$ are not profitable, since the entrant would obtain the same demand as in equilibrium, but lower engagement. If the entrant deviates by setting $h_E \in [0, \tilde{h}_E]$, the profits it obtains are bounded from above by $\rho\pi_E^r(e_E^*(h_E))$. This is because the entrant is not joined by naive users if it deviates in this way, but is joined by rational users. Thus, the entrant has no profitable deviations if and only if:

$$\rho\pi_E^r(e_E^*(\tilde{h}_E)) \leq (1-\rho)\pi_E^n(e_E^*(1)) \iff \frac{\rho}{1-\rho} \leq \frac{\pi_E^n(e_E^*(1))}{\pi_E^r(e_E^*(\tilde{h}_E))}$$

Secondly, consider profitable deviations by the incumbent. In equilibrium, it is joined by rational users and obtains the profits $\rho\pi_I^r(e_I^*(\tilde{h}_I))$. All deviations $h_I \in [0, \tilde{h}_I)$ cannot be

profitable, since these leave demand unaffected and reduce engagement. When the incumbent deviates by setting any $h_I \in (\tilde{h}_I, 1]$, it obtains no demand from rational users. Thus, the most profitable deviation would be to $h_I = 1$. If the incumbent sets $h_I = 1$, it is joined by all naive users and obtains deviation profits of $(1-\rho)\pi_I^n(e_I^*(1))$. Thus, the incumbent has no profitable deviations if and only if:

$$(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\tilde{h}_I)) \iff \frac{\pi_I^n(e_I^*(1))}{\pi_I^r(e_I^*(\tilde{h}_I))} \leq \frac{\rho}{1-\rho}$$

In summation, the platforms have no profitable deviations if and only if:

$$\frac{\pi_I^n(e_I^*(1))}{\pi_I^r(e_I^*(\tilde{h}_I))} \leq \frac{\rho}{1-\rho} \leq \frac{\pi_E^n(e_E^*(1))}{\pi_E^r(e_E^*(\tilde{h}_E))}$$

**Equilibrium candidate 3:** An equilibrium in which $h_E^* = 0$ and $h_I^* = \check{h}_I$ exists if and only if the following conditions hold jointly: (i) $U_E^n(1) \leq U_I^n(h_I^*)$ and (ii) $(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$.

In equilibrium, rational users are indifferent between joining the entrant and the incumbent, so it is optimal for them to all join the incumbent. Moreover, $h_I^* \geq 0 = h_E^*$ must hold, which implies that naive users will strictly prefer to join the incumbent in equilibrium.

Firstly, consider possible deviations for the entrant. The most profitable deviation for the entrant is to set $h_E = 1$. For any $h_E \in (0, 1]$, rational users will not join the entrant since they are indifferent in equilibrium and because $U_E^r(h_E)$ is strictly decreasing. This directly implies that the most profitable deviation would be to set $h_E = 1$, because this deviation maximizes naive users' utility of joining the entrant (and thus the demand the entrant obtains) as well as the engagement the entrant obtains.

The deviation $h_E = 1$ would be profitable for the entrant if and only if $U_E^n(1) > U_I^n(h_I^*)$. This holds because naive users strictly prefer to join the entrant after the deviation if $U_E^n(1) > U_I^n(h_I^*)$, in which case the entrant obtains strictly positive profits when setting $h_E = 1$.

Second, consider possible deviations for the incumbent. Any deviation below $h_I^*$ cannot be profitable, since this cannot increase its demand (all users already join the incumbent in equilibrium), but will reduce engagement. If the incumbent deviates by increasing $h_I$, it will not be joined by rational users anymore. Thus, the most profitable deviation is to $h_I = 1$, since this guarantees that it is joined by naive users and maximizes engagement. The equilibrium profits are $\rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$, while the deviation profits are

$(1-\rho)\pi_I^n(e_I^*(1))$. Thus, the deviation is not profitable if and only if:

$$(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I)) \tag{16}$$

■

**Proof of Corollary 1:**

**Part 1:** Existence.

By previous arguments, an equilibrium in which $h_I^* = 1$ and $h_E = \tilde{h}_E$ exists if:

$$\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \leq (1-\rho)\pi_I^n(e_I^*(1)) \tag{17}$$

Note that the right-hand side of this equality is strictly decreasing in $\rho$ and equals $\pi_I^n(e_I^*(1))$ if $\rho = 0$. At $\rho = 0$, the right-hand side is thus strictly larger than the left-hand side (since $e_I^*(\tilde{h}_I) < e_I^*(1)$ holds because $\tilde{h}_I < 1$). Define $\underline{\rho}$ such that $\underline{\rho}\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\underline{\rho})\pi_I^n(e_I^*(\tilde{h}_I)) = (1-\underline{\rho})\pi_I^n(e_I^*(1))$. For all $\rho \in (0, \underline{\rho})$, the equilibrium thus exists.

**Part 2:** Uniqueness.

Consider any $\rho < \underline{\rho}$, which implies that $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) < (1-\rho)\pi_I^n(e_I^*(1))$.

Suppose, for a contradiction, that there exists a pure-strategy equilibrium in which $h_I^* < 1$. Then, $h_I^* \leq \tilde{h}_I$ must hold. The profits which the incumbent makes in equilibrium are thus bounded from above by $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$. Since $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))) < (1-\rho)\pi_I^n(e_I^*(1))$, the incumbent would strictly prefer to deviate from the equilibrium by setting $h_I = 1$.[21] This is a contradiction.

There exists no mixed-strategy equilibrium under the stated condition. Suppose, for a contradiction, that there exists an equilibrium in which the incumbent mixes. When setting any $h_I < 1$, the incumbent's profits are strictly below $(1-\rho)\pi_I^n(e_I^*(1))$, i.e., the incumbent's profits when it sets $h_I = 1$. Thus, the mixing indifference condition cannot hold.

Thus, the incumbent cannot mix in equilibrium and will set $h_I^* = 1$. But then, the entrant's profits attain a strict maximum at $h_E = \tilde{h}_E$. Thus, the entrant would also mix. ■

[21] This holds because the incumbent will be joined by all naive users if it sets $h_I = 1$, given that it has a competitive advantage.

**Proof of Lemma 2:**

Suppose, for a contradiction, that there exists a mixed-strategy equilibrium (MSE) in which all users join the entrant with probability 1. Then, the incumbent obtains zero profits in equilibrium. But then, the incumbent would strictly prefer to deviate by setting $h_I = 1$. After the deviation, all naive users join the entrant and choose positive engagement, which means that the incumbent obtains positive profits through the deviation. This is a contradiction, which means that such an equilibrium does not exist.

Now consider an equilibrium in which all users join the incumbent with probability 1. Then, the incumbent will set a given harmful content share with probability 1. Suppose, for a contradiction, that there exist two different harmful content levels in the support of $\Gamma_I$. Since the incumbent is joined by all users with probability 1, the demand which the incumbent obtains for all harmful content shares it offers on the equilibrium path must be the same, but one harmful content share must yield strictly higher engagement from all users (by Assumption 1), and thus, strictly higher profits. This is a contradiction. Define the harmful content share the incumbent sets as $h_I^*$.

Suppose, for a contradiction, that there exists an equilibrium in which $U_I^r(h_I^*) > U_E^r(0)$. Then, all rational users strictly prefer to join the incumbent, no matter the harmful content share the entrant sets. When the incumbent marginally increases $h_I$, all rational users still strictly prefer to join the incumbent, no matter what $h_E$ the entrant sets. The marginal increase of $h_I$ also weakly increases the demand the incumbent receives from naive users and strictly increases engagement. This makes the deviation profitable, a contradiction.

Suppose, for a contradiction, that there exists an equilibrium in which $U_I^r(h_I^*) < U_E^r(0)$. Then, the entrant could set a harmful content level just above $h_E = 0$ to obtain positive profits (since rational users then strictly prefer to join the entrant). Thus, the entrant would have a profitable deviation, since it obtains zero profits in equilibrium, a contradiction.

Thus, $U_I^r(h_I^*) = U_E^r(0)$ must hold in equilibrium. This equation has a unique solution, namely $h_I^* = \check{h}_I$. ■

**Proof of Proposition 4:**

Consider any MSE in which some users join the entrant with positive probability. Define $\check{h}_I$ such that $U_I^r(\check{h}_I) = U_E^r(0) := \bar{U}$, where $\bar{U}$ is the utility which all users obtain in the pure-strategy equilibrium in which all users join the incumbent (or in any mixed-strategy equilibrium in which all users join the incumbent with probability 1, by the arguments made

in Lemma 2).

By definition, the entrant must set a harmful content share weakly above $h_E = 0$. The incumbent would never set a harmful content level strictly below $\check{h}_I$. To see this, note that $\check{h}_I < \tilde{h}_I$. For any $h_I < \check{h}_I$, rational users would thus strictly prefer to join the incumbent for any $h_E$ set by the entrant (since $U_I^r(h_I) > U_I^r(\check{h}_I) = U_E^r(0) > U_E^r(h_E)$ holds for any $h_I < \check{h}_I$ and any $h_E \in [0,1]$). For any $h_I < \check{h}_I$, the incumbent would thus strictly prefer to set a harmful content share just above this $h_I$ (since this yields higher engagement).

In a mixed-strategy equilibrium, the incumbent must set a harmful content level $h_I > \check{h}_I$ with strictly positive probability or the entrant must set a harmful content level $h_E > 0$ with strictly positive probability (by definition, otherwise both firms would not be mixing).

For any $h_E > 0$ such that $h_E \in \text{supp}\Gamma_E$, the ordering $U_E^r(h_E) < \bar{U}$ must hold. For any $h_I > \check{h}_I$ such that $h_I \in \text{supp}\Gamma_I$, the inequality $U_I^r(h_I) < \bar{U}$ must hold. Both statements hold because $U_p^r(h_p)$ is strictly decreasing in $h_p$ for either $p \in \{I, E\}$ by Assumption 1.

Suppose the incumbent sets a harmful content level strictly above $\check{h}_I$ with strictly positive probability. For any $h_I > \check{h}_I$ such that $h_I \in \text{supp}\Gamma_I$, the demand which the incumbent obtains must be strictly positive (i.e. some users must join the incumbent and obtain utility $U_I^r(h_I) < \bar{U}$).[22] Thus, users will receive a utility strictly below $\bar{U}$ with strictly positive probability, which implies the result.

Suppose the entrant sets a harmful content level strictly above 0 with strictly positive probability, and the incumbent sets the harmful content level $\check{h}_I$ with probability 1. For any $h_E > 0$ such that $h_E \in \text{supp}\Gamma_E$, the demand which the incumbent obtains must be strictly positive. Otherwise, we are outside of the space of equilibria we consider. Thus, the probability that some user type joins the entrant and obtains a utility level strictly below $\bar{U}$ is strictly positive, which implies the desired result. ■

**Proof of Proposition 5:**

**Part 1:** A unique mixed-strategy equilibrium exists if $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $(1-\rho)e_I^*(1) \in \left(e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\right)$.

Assume that $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $(1-\rho)e_I^*(1) \in \left[e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\right]$. Lemma 8, which we present in the Online Appendix, establishes that any mixed-strategy equilibrium must have the structure we described in the proposition.

[22] If the incumbent obtains zero demand (and thus obtains zero profits) when setting some $h_I \in \text{supp}\Gamma_I$, it would have a profitable deviation, since it can always obtain strictly positive profits by setting $h_I < \check{h}_I$.

A mixed-strategy equilibrium (MSE) with the described properties exists if and only if there exist $\lambda_I$, $\lambda_E$, $\underline{h}_I$, and $\underline{h}_E$ that constitute joint solutions to the following set of equations:

$$\Pi_I(1) = \lim_{h_I \uparrow \tilde{h}_I} \Pi_I(h_I) \iff e_I^*(1)[1-\rho] = e_I^*(\tilde{h}_I)\big[\rho\lambda_E + (1-\rho)\big] \tag{18}$$

$$\Pi_I(1) = \Pi_I(\underline{h}_I) \iff (1-\rho)e_I^*(1) = e_I^*(\underline{h}_I) \tag{19}$$

$$\Pi_E(\underline{h}_E) = \Pi_E(\tilde{h}_I) \iff e_E^*(\underline{h}_E)\big[\rho\big] = e_E^*(\tilde{h}_E)\big[\rho\lambda_I\big] \tag{20}$$

$$U_E^r(\underline{h}_E) = U_I^r(\underline{h}_I) \tag{21}$$

Note that $\lambda_I$ is the probability that the incumbent plays $h_I = 1$, and $\lambda_E$ is the probability that the entrant plays $h_E = \tilde{h}_E$.

(i) Showing that a joint solution to equations (18) - (20) exists.

Firstly, note that there exists (under our assumptions) a unique $\lambda_E \in [0,1]$ that solves equation (18). To see this, note that left hand side of equation (18) is larger than the right hand side if $\lambda_E = 0$. In addition, the left hand side of equation (18) is smaller than the right hand side if $\lambda_E = 1$ because $(1-\rho)e_I^*(1) \leq e_I^*(\tilde{h}_I)$ holds by assumption. The fact that the right-hand side of equation (18) is continuous and strictly increasing in $\lambda_E$ then implies the desired result.

Secondly, note that there always exists a unique $\lambda_I$ such that equation (20) is satisfied.

Thirdly, there exists (under our assumptions) a unique $\underline{h}_I \in (\check{h}_I, \tilde{h}_I)$ that solves equation (19). To see this, note that the left hand side of this equation is larger than the right hand side if $\underline{h}_I = \check{h}_I$, and that the left hand side of this equation is smaller than the right hand side if $\underline{h}_I = \tilde{h}_I$. The fact that the right-hand side of equation (19) is continuous and increasing in $\underline{h}_I$ then implies the desired result.

Fourthly, note that a unique solution $\underline{h}_E$ of equation (21) exists if $\underline{h}_I \geq \check{h}_I$ holds (we have verified the existence of an appropriate $\underline{h}_I$ in the last term. To see this, note that $U_I^n(\underline{h}_I) \geq U_I^n(\check{h}_I)$ holds because $\underline{h}_I \geq \check{h}_I$. If $\underline{h}_E = 0$, the left-hand side of equation (21) is thus weakly larger than the right-hand side. If $\underline{h}_E = \tilde{h}_E$, the left-hand side of equation (21) is strictly smaller than the right-hand side. The fact that the left-hand side of (21) is continuous and strictly decreasing in $\underline{h}_E$ then implies the desired result.

(ii) Since a joint solution $(\underline{h}_I, \underline{h}_E, \lambda_I, \lambda_E)$ of equations (18) - (20) exists, a mixed-strategy

equilibrium exists.

To show this, we first set $F_I(h_I)$ on $h_I \in [\underline{h}_I, \tilde{h}_I]$ and $F_E(h_E)$ on $h_E \in [\underline{h}_E, \tilde{h}_E]$ appropriately, and then establish that there are no profitable deviations.

To do this, we define a function $r(h_I)$ such that, if the incumbent sets $h_I$ and the entrant sets $h_E$, rational users join the entrant if $h_E < r(h_I)$. Note that this function is increasing. For any $h_E$, the profits the entrant obtains are given by:

$$e_E^*(h_E)\rho\big[1 - F_I(r^{-1}(h_E))\big]$$

For any such $h_E$, find the $h_I \in [\underline{h}_I, \tilde{h}_I]$ such that $h_I = r^{-1}(h_E)$, i.e. $h_E = r(h_I)$. Thus, any $h_I \in [\underline{h}_I, \tilde{h}_I]$ needs to solve:

$$e_E^*(r(h_I))\rho\big[1 - F_I(h_I)\big] = e_E^*(\underline{h}_E)\rho, \tag{22}$$

where the profits the entrant obtains when setting $\underline{h}_E$ are given by the right-hand side. Thus, the value of $F_I(h_I)$ must satisfy:

$$F_I(h_I) = 1 - \frac{e_E^*(\underline{h}_E)}{e_E^*(r(h_I))}$$

Now we pin down $F_E(h_E)$ for any $h_E \in (\underline{h}_E, \tilde{h}_E)$ by considering the incumbent's profits. For any $h_I \in (\underline{h}_I, \tilde{h}_I)$, the profits the incumbent obtains are given by:

$$e_I^*(h_I)\big[(1-\rho) + \rho\big(1 - F_E(r(h_I))\big)\big]$$

For any such $h_I$, we can find $h_E \in (\underline{h}_E, \tilde{h}_E)$ such that $r(h_I) = h_E$. For any $h_E$, we thus need to have:

$$e_I^*(r^{-1}(h_E))\big[(1-\rho) + \rho\big(1 - F_E(h_E)\big)\big] = e_I^*(\underline{h}_I) \tag{23}$$

Solving for $F_E(h_E)$ yields:

$$(1-\rho) + \rho\big(1 - F_E(h_E)\big) = \frac{e_I^*(\underline{h}_I)}{e_I^*(r^{-1}(h_E))} \iff 1 - F_E(h_E) = -\frac{1-\rho}{\rho} + \frac{e_I^*(\underline{h}_I)}{\rho e_I^*(r^{-1}(h_E))}$$

$$\iff$$

$$F_E(h_E) = \frac{1}{\rho} - \frac{e_I^*(\underline{h}_I)}{\rho e_I^*(r^{-1}(h_E))} \tag{24}$$

Then, no platform will have any profitable deviations. For any $p$, all $h_p \in [\underline{h}_p, \tilde{h}_p)$ yield the same profits by construction. All $h_p < \underline{h}_p$ yield lower profits than $\underline{h}_p$. For the entrant, any $h_E \in (\tilde{h}_E, 1]$ yields lower profits by assumption. For the entrant, all $h_I \in (\tilde{h}_I, 1)$ yield lower profits than $h_I = 1$ (which, in turn, yields the same profits as any $h_I \in \text{supp}\Gamma_I$ by construction).

**Part 2:** A mixed-strategy equilibrium does not exist if $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $(1-\rho)e_I^*(1) \notin \big(e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\big)$.

Suppose firstly that $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $(1-\rho)e_I^*(1) < e_I^*(\check{h}_I)$. Then, the incument would never set the harm content share 1. Thus, no mixed-strategy equilibrium can exist because $1 \in \text{supp}\Gamma_I$ must hold in any mixed-strategy equilibrium if $U_E^n(1) \leq U_I^n(\check{h}_I)$ (see Lemma 8).

Suppose secondly that $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $(1-\rho)e_I^*(1) = e_I^*(\check{h}_I)$. Lemmas 6 - 8 pin down that, if a mixed-strategy equilibrium exists, then the entrant must draw harmful content shares from a distribution with gapless support on $[\underline{h}_E, \tilde{h}_E]$.

Since $(1-\rho)e_I^*(1) = e_I^*(\check{h}_I)$, $\underline{h}_I = \check{h}_I$ must hold in equilibrium. This implies that $\underline{h}_E = 0$. But then, the entrant's mixing indifference condition cannot hold, a contradiction. Hence, there exists no mixed-strategy equilibrium.

Suppose secondly that $U_E^n(1) \leq U_I^n(\check{h}_I)$ and $e_I^*(\tilde{h}_I) \leq (1-\rho)e_I^*(1)$. Then, the incumbent's mixing indifference condition cannot be satisfied because $\lim_{h_I \to \tilde{h}_I} \Pi_I(h_I) < e_I^*(\tilde{h}_I) \leq (1-\rho)e_I^*(1) = \Pi_I(1)$. This is a contradiction. ■

**Proof of Proposition 6:** We show that there exists a unique equilibrium for any parameter combination at which $U_E^n(1) \leq U_I^n(\check{h}_I)$.

Suppose $(1-\rho)e_I^*(1) \leq e_I^*(\check{h}_I) < e_I^*(\tilde{h}_I)$. Then, the equilibrium in which $(h_I^*, h_E^*) = (\check{h}_I, 0)$ exists. There exists no mixed-strategy equilibrium by Proposition 5, and there exists no other pure-strategy equilibrium by Proposition 3: An equilibrium in which $(h_I^*, h_E^*) = (1, \tilde{h}_E)$ does not exist because $(1-\rho)e_I^*(1) < e_I^*(\tilde{h}_I)$, and an equilibrium in which $(h_I^*, h_E^*) = (\tilde{h}_I, 1)$ does not exist because $U_E^n(1) \leq U_I^n(\check{h}_I)$, which implies that $U_E^n(1) < U_I^n(\tilde{h}_I)$. Thus, there exists a unique equilibrium.

Suppose $(1-\rho)e_I^*(1) \in \big(e_I^*(\check{h}_I), e_I^*(\tilde{h}_I)\big)$. By Proposition 3 and previous arguments, there exists no pure-strategy equilibrium. By Proposition 5, there exists a mixed-strategy equilibrium, which is hence unique.

Suppose $e_I^*(\check{h}_I) < e_I^*(\tilde{h}_I) \leq (1-\rho)e_I^*(1)$. By previous arguments, there exists no mixed-strategy equilibrium, and a unique pure-strategy equilibrium, namely the equilibrium in

which $(h_I^*, h_E^*) = (1, \tilde{h}_E)$. ■

**Proof of Proposition 7:**

**Part 1:** If $(1-\rho)e_I^*(1) \leq 0.5e_I^*(0)$, there exists an equilibrium in which $h_I^* = h_E^* = 0$.

The proof is by construction. Note that $e_I^*(h) = e_E^*(h)$ holds for both $p \in \{E, I\}$ because there is no competitive advantage.

Suppose $h_I^* = h_E^* = 0$. Then, it is optimal for all users to join both platforms with probability 0.5. In equilibrium, the profits of either platform are $0.5e_I^*(0)$.

Now consider profitable deviations by any platform $p$. For any deviation $h_p > 0$, the platform $p$ is not joined by any rational users, but by all naive users. Thus, the most profitable deviation of any platform is to set a harmful content share equal to 1.

When deviating by setting $h_p = 1$, the profits which the platform $p$ obtains are given by $(1-\rho)e_I^*(1)$. Since there is no technological gap, $e_I^*(1) = e_E^*(1)$ holds. Hence, there are no profitable deviations for either platform if and only if the stated condition holds.

**Part 2:** If $(1-\rho)e_I^*(1) \leq 0.5e_I^*(0)$, there exists a unique pure-strategy equilibrium.

To begin, note the following: If $(1-\rho)e_I^*(1) \leq 0.5e_I^*(0)$, then $\rho > 0.5$ must hold.

There exists no pure-strategy equilibrium in which $h_I^* = h_E^*$. Suppose, for a contradiction, that there exists an equilibrium in which $h_I^* = h_E^* := h^* > 0$. Then, some platform $p$ would strictly prefer to set a harmful content share marginally below $h_p$, because this guarantees that all rational users join platform $p$. Thus, the deviation induces an upward jump in the demand which the platform obtains (since $\rho > 0.5$), while changing the optimal engagement level of the platform's users in continuous fashion. This is a contradiction.

There exists no pure-strategy equilibrium in which $h_I^* < h_E^*$. Suppose, for a contradiction, that such an equilibrium exists. In such a hypothetical equilibrium, rational users would strictly prefer to join the incumbent. Thus, $h_E^* = 1$ must hold. But this implies that $h_I^* = \tilde{h}_I$ must hold — otherwise, the incumbent would strictly prefer to increase $h_I$. But then, the entrant would prefer to deviate by setting $h_E = 0$, given that all rational users would join the entrant after the deviation and $(1-\rho)e_I^*(1) \leq 0.5e_I^*(0)$.

Analogous arguments imply that there exists no equilibrium in which $h_E^* < h_I^*$. ■

**Proof of Proposition 8:**

**Part 1:** In any equilibrium in which some users multi-home, the users which multi-home must choose zero engagement on one platform.

Suppose, for a contradiction, that there exists an equilibrium in which some user (no matter whether she is naive or rational) multi-homes and devotes positive engagement on both platforms. Recall that there are no network effects.

If the user multi-homes, her engagement choices must maximize the following object:

$$U^B(e_I, e_E, h_I, h_E) = \big(\eta_I h_I e_I + \theta_I(1-h_I)e_I\big) + \big(\eta_E h_E e_E + \theta_E(1-h_E)e_E\big) - \gamma\big(e_I + e_E\big)^2$$

If the user devotes positive engagement on both platforms, the following must hold:

$$\frac{\partial U^B}{\partial e_I} = 0 = \frac{\partial U^B}{\partial e_E} \iff \eta_I h_I + \theta_I(1-h_I) = 2\gamma\big(e_I + e_E\big) = \eta_E h_E + \theta_E(1-h_E) \tag{25}$$

$$\iff$$

$$\eta_I h_I + \theta_I(1-h_I) = \eta_E h_E + \theta_E(1-h_E) \tag{26}$$

In turn, this implies that the entrant must have a larger share of harmful content, i.e. that $h_E > h_I$. Else, these two objects could not be equal. To see this, note that $\eta_E h_E + \theta_E(1-h_E)$ is increasing in $h_E$ and, if $h_E = h_I$, the left-hand side of this equation would be larger because of the competitive advantage. Thus, $h_E > h_I$ must hold.

We refer to the total engagement level $e_I + e_E$ that solves the first-order condition in equation (25) as $\bar{e}^*(h)$.

The total utility a rational user obtains if she just joins the incumbent and devotes engagement $\bar{e}^*(h)$ there is given by $\big(\eta_I h_I + \theta_I(1-h_I)\big)\bar{e}^*(h) + g_I - \delta h_I - \gamma\big(\bar{e}^*(h)\big)^2$.

The total utility the user attains through multi-homing is given by:

$$\big(\eta_I h_I + \theta_I(1-h_I)\big)\bar{e}^*(h) + 0.5(g_I - \delta h_I) + 0.5(g_E - \delta h_E) - \gamma\big(\bar{e}^*(h)\big)^2$$

This is because $\eta_I h_I + \theta_I(1-h_I) = \eta_E h_E + \theta_E(1-h_E)$ must hold in the postulated equilibrium.

Given that $h_E > h_I$ holds, a rational user would thus attain larger utility by just joining the entrant and devoting the engagement level $\bar{e}^*(h)$ on the incumbent platform. This is because the following inequality holds (given that $g_I > g_E$ and $-\delta h_I > -\delta h_E$):

$$\big(\eta_I h_I + \theta_I(1-h_I)\big)\bar{e}^*(h) + g_I - \delta h_I - \gamma\big(\bar{e}^*(h)\big)^2 >$$

$$\big(\eta_I h_I + \theta_I(1-h_I)\big)\bar{e}^*(h) + 0.5(g_I - \delta h_I) + 0.5(g_E - \delta h_E) - \gamma\big(\bar{e}^*(h)\big)^2$$

Taken together, the previous arguments establish that there exists no equilibrium in which rational users multi-home and choose positive engagement levels on both platforms.

Analogous arguments establish that there also exists no equilibrium in which naive users multi-home and choose positive engagement levels on both platforms. In such an equilibrium, naive users must be indifferent between engagement on both platforms. But then, naive users would prefer to only join the incumbent, a contradiction.

**Part 2:** In any equilibrium in which some users multi-home and choose positive engagement on the entrant, the utility of rational users must be zero and the utility of naive users must be strictly negative.

We establish this result in three steps. We begin by showing that (i) there exists no equilibrium in which all users multi-home and that (ii) there exists no equilibrium in which only rational users multi-home. Finally, we show that (iii) in any equilibrium in which naive users multi-home, rational users must obtain zero utility and naive users must obtain weakly negative utility.

(i) There exists no equilibrium in which all users multi-home (and some users choose positive engagement on the entrant platform).

Suppose, for a contradiction, that such an equilibrium exists. If all users choose zero engagement on the entrant's platform, we are outside of the equilibria we consider, a contradiction. If all users choose zero engagement on the incumbent's platform, the incumbent obtains zero profits in equilibrium. But then, the incumbent would prefer to deviate by setting $h_I = 1$, a contradiction. Finally, consider an equilibrium in which rational users choose zero engagement on one platform and naive users choose zero engagement on the other platform. In order for such an equilibrium to exist, all users must be indifferent between spending time on the incumbent platform and the entrant platform. By previous arguments, this implies that $h_I^* > h_E^*$ must hold, which means that all users would strictly prefer to join the incumbent instead of multi-homing, a contradiction.

(ii) There exists no equilibrium in which only rational users multi-home.

Suppose rational users multi-home and devote zero engagement on the incumbent platform. Then, naive users must join the incumbent (else, the incumbent would deviate). Because the incumbent only obtains profits from naive users in equilibrium, it would optimally

set $h_I = 1$. But then, rational users would not multi-home, a contradiction.

Suppose rational users multi-home and devote zero engagement on the entrant platform. Then, naive users must join the entrant (else, we are outside of the space of equilibria we consider). Since the entrant only obtains profits from naive users, it will set $h_E = 1$. But then, rational users would not multi-home, a contradiction.

(iii) In any equilibrium in which naive users multi-home, rational users must obtain zero utility and naive users must obtain weakly negative utility.

First, suppose naive users multi-home and devote zero engagement on the entrant platform. Then, rational users must join the entrant (else, we are outside of the space of equilibria we consider). Hence, the incumbent only obtains profits from naive users.

This implies that $h_I = 1$ must hold. If $h_I < 1$, the incumbent would prefer to raise $h_I$ to increase the engagement it receives from its users. But then, rational users must obtain zero utility by joining the entrant. Suppose, for a contradiction, that they obtain positive utility. Then, they strictly prefer to join the entrant. Hence, the entrant would prefer to slightly increase $h_E$, since this leaves its demand unaffected, but increases engagement.

Thus, all users who join the entrant obtain zero utility in equilibrium. Moreover, the fact that $h_I = 1$ implies that naive users who join the incumbent must attain negative utility.

Second, suppose alternatively that naive users multi-home and devote zero engagement on the incumbent platform. Then, rational users must join the incumbent (else, the incumbent would prefer to deviate). Hence, the entrant only derives profits from naive users. By previous arguments, this implies that $h_E = 1$ must hold. This establishes that rational users must obtain zero utility by joining the incumbent in equilibrium (else, it would prefer to raise $h_I$). By implication, all users who join the incumbent obtain zero utility and all users who join the entrant obtain negative utility. ■

**Proof of Proposition 9:**

**Part 1:** In any equilibrium in which the entrant is joined by some users, all rational users obtain utility zero and all naive users obtain utility weakly below $U_I(1, e_I^n(1))$.

Firstly, consider an equilibrium in which both platforms and all users play a pure strategy. Suppose all rational users join platform $p$ and all naive users join platform $l \neq p$.

In equilibrium, platform $l$ must set $h_l = 1$. Suppose, for a contradiction, that $h_l < 1$. In equilibrium, naive users must weakly prefer this platform, and rational users join the other platform. The participation constraint of naive users is always slack. If platform $l$ deviates

by setting $h_l = 1$, this will raise the utility that naive users attain on platform $l$, so they would still choose to join this platform after the deviation. Moreover, rational users do not join platform $l$ in equilibrium. Thus, the deviation raises the total engagement that platform $l$ receives without reducing its demand. Hence, the deviation is profitable, a contradiction.

The fact that $h_l = 1$ must hold means that rational users would attain negative utility when joining platform $l$ (by Assumption 1). It also implies that naive users obtain utility below $U_I(1, e_I^n(1))$ in equilibrium by Assumption 1.

In equilibrium, rational users must obtain zero utility. Suppose, for a contradiction, that rational users attain strictly positive utility by joining platform $p$. Then, platform $p$ would find it optimal to marginally increase the share of harmful content it displays. After the deviation, rational users would still strictly prefer to join platform $p$ (since rational users would obtain negative utility by joining platform $l$), but the platform obtains higher engagement from all rational users who join it. If naive users would also join the platform after the deviation, the deviation become even more profitable. Hence, the deviation is profitable, which is a contradiction.

Secondly, consider equilibria in which both platforms play a pure strategy and some users play a mixed strategy. Suppose naive users mix (which means they must be indifferent between both platforms). This means that rational users cannot mix.[23] Suppose rational users join platform $p$. This means that platform $l$ is only joined by naive users, so it will optimally set $h_l^* = 1$. By implication, this implies that $h_p^* = \tilde{h}_p$ must hold. All users who join platform $p$ obtain zero utility in equilibrium, while all users who join platform $l$ obtain utility below $U_I(1, e_I^n(1))$ in equilibrium.

Finally, note that there exists no equilibrium in which platforms play a pure strategy and rational users mix. In such an equilibrium, naive users must strictly prefer to join some platform. Suppose naive users join platform $p$. Then, platform $l$ would strictly prefer to marginally reduce the harmful content share it offers, because all rational users join platform $l$ after the deviation.

**Part 2:** In any equilibrium in which all users join the incumbent, all users rational users obtain strictly positive utility and all naive users obtain utility strictly above $U_I(1, e_I^n(1))$.

Note that the entrant can always guarantee that any rational user who joins it obtains positive utility by setting a harmful content share in a small open interval above zero. If the entrant sets such a harmful content share and is joined by rational users, it obtains strictly

[23] Suppose naive users are indifferent. Because the incumbent has a competitive advantage, $h_I < h_E$ must hold. This means that rational users strictly prefer to join the incumbent.

positive profits.

Suppose, for a contradiction, that rational users joins the incumbent but attain utility zero. Then, the entrant would deviate by setting a harmful content share in a small open interval above zero. After the deviation, rational users would join the entrant and choose positive engagement. Thus, the deviation is profitable because it enables the entrant to obtain positive profits (while it obtains zero profits in equilibrium). This is a contradiction.

Hence, rational users obtain strictly positive utility in equilibrium when joining the incumbent. This implies that $h_I^* < 1$ must hold.

The utility which naive users obtain in equilibrium is $U_I(h_I^*, e_I^n(h_I^*))$. By assumption 1, the function $U_I(h, e_I^n(h))$ is falling in $h$. This implies that $U_I(h_I^*, e_I^n(h_I^*)) > U_I(1, e_I^n(1))$. ■

**Proof of Proposition 10:**

**Part 1:** In any equilibrium in which the entrant is joined by some users, rational users obtain zero utility and naive users obtain negative utility.

Firstly, consider an equilibrium in which both platforms and all users play a pure strategy. Suppose all rational users join platform $p$ and all naive users join platform $l \neq p$.

In equilibrium, platform $l$ must set $h_l = 1$. The fact that $h_l = 1$ must hold means that rational users would attain negative utility when joining platform $l$ (by Assumption 2). It also implies that naive users obtain negative utility in equilibrium.

In equilibrium, rational users must obtain zero utility. Suppose, for a contradiction, that rational users attain strictly positive utility by joining platform $p$. Then, platform $p$ would find it optimal to marginally increase the share of harmful content it displays, since this can only weakly raise the demand it obtains (which could only further benefit it through network effects) and will increase the engagement of all users on its platform.

Secondly, consider equilibria in which both platforms play a pure strategy and some users play a mixed strategy. Suppose naive users mix (which means they must be indifferent between both platforms). This means that rational users cannot mix.[24] Suppose rational users join platform $p$. This means that platform $l$ is only joined by naive users, so it will optimally set $h_l^* = 1$. By implication, this implies that $h_p^* = \tilde{h}_p$ must hold. All users who join platform $p$ obtain zero utility in equilibrium, while all users who join platform $l$ obtain negative utility in equilibrium.

[24]Suppose naive users are indifferent. Because the incumbent has a competitive advantage and the incumbent's network size is larger, $h_I < h_E$ must hold. This means that rational users strictly prefer to join the incumbent.

Finally, note that there exists no equilibrium in which platforms play a pure strategy and rational users mix. In such an equilibrium, naive users must strictly prefer to join some platform. Suppose naive users join platform $p$. Then, platform $l$ would strictly prefer to marginally reduce the harmful content share it offers, because all rational users join platform $l$ after the deviation.

**Part 2:** In any equilibrium in which all users join the incumbent, all users users obtain strictly positive utility

Note that the entrant can always guarantee that any rational user who joins it obtains positive utility by setting a harmful content share in a small open interval above zero. This is because $U_p^r(0, s_p) > 0$ holds for any $s_p$ (by Assumption 2). If the entrant sets such a harmful content share and is joined by rational users, it obtains strictly positive profits.

Suppose, for a contradiction, that rational users joins the incumbent but attain utility zero. Then, the entrant would deviate by setting a harmful content share in a small open interval above zero. After the deviation, rational users would join the entrant and choose positive engagement. Thus, the deviation is profitable because it enables the entrant to obtain positive profits (while it obtains zero profits in equilibrium). This is a contradiction.

Hence, rational users must obtain positive utility in equilibrium. Since rational users' and naive users' chosen engagement levels are given by the same function $e_p^*(h_p, s_p)$, naive users also obtain positive utility in equilibrium. ■

**Proof of Proposition 11:**

**Part 1:** In any pure-strategy equilibrium in which all users join the incumbent, all users obtain strictly positive utility. In any pure-strategy equilibrium in which some users join the entrant, all users obtain weakly negative utility.

Consider an equilibrium in which all users (including rational users) join the incumbent. Suppose, for a contradiction, that users obtain weakly negative utility when joining the incumbent. Then, the entrant would prefer to deviate from the equilibrium by setting $h_E$ in an open interval above 0 to attract all rational users, since this enables them to obtain strictly positive profits in equilibrium. This is a contradiction.

Now consider an equilibrium where some users join the entrant. These users must be rational and cannot be indifferent in equilibrium (else, the incumbent would prefer to slightly reduce the share of harmful content it shows to attract all users). Thus, all rational users join

the entrant in the postulated equilibrium. Hence, the incumbent is only joined by captive users, which implies that $h_I^* = 1$ must hold. Then, the entrant would optimally set $h_E^* = \tilde{h}_E$, given that rational users would obtain negative utility by joining the entrant.

**Part 2:** There exists a $\rho^1$ s.t., if $\rho < \rho^1$, there is a unique equilibrium in which $h_I^* = 1$, $h_E^* = \tilde{h}_E$, rational users join the entrant and the market share of the incumbent is $1 - \rho$.

We define $\rho^1$ such that $(1 - \rho^1)e_I^*(1) = e_I^*(\tilde{h}_I)$. If $\rho < \rho^1$, then $(1 - \rho^1)e_I^*(1) > e_I^*(\tilde{h}_I)$. Thus, the equilibrium in which $h_I^* = 1$ and $h_E^* = \tilde{h}_E$ exists and is unique. This is because the entrant has no profitable deviations and the incumbent's most profitable deviation would be to $h_I = \tilde{h}_I$, which is not profitable under the stated condition. The equilibrium is unique because the incumbent would never find it optimal to set a harmful content share below 1.

**Part 3:** There exists a $\rho^2$ s.t., if $\rho > \rho^2$, there is a unique equilibrium in which $h_I^* = \tilde{h}_I$ and all users join the incumbent.

Now, we define $\rho^2$ such that $(1 - \rho^2)e_I^*(1) = e_I^*(\check{h}_I)$ If $\rho > \rho^2$, then $(1 - \rho^2)e_I^*(1) < e_I^*(\check{h}_I)$. Thus, there exists a unique equilibrium in which $h_I^* = \check{h}_I$. Existence of this equilibrium follows from the fact that the most profitable deviation for the incumbent (namely, setting $h_I = 1$) is not profitable under the stated condition. The entrant cannot attract any users because $U_E^r(h_E) \leq U_I^r(\check{h}_I)$ holds for all $h_E \in [0, 1]$.

We now establish uniqueness of this equilibrium. Firstly, note that there exists no other pure-strategy equilibrium, because it is never optimal for the incumbent to set $h_I^* = 1$.

Secondly, note that there exists no mixed-strategy equilibrium in which the entrant is joined by a positive measure of users under the stated condition. In any such mixed-strategy equilibrium (where we label the distributions of harmful content shares for the incumbent and the entrant $\Gamma_I$ and $\Gamma_E$, respectively), the entrant would only ever set harmful content shares below $\tilde{h}_E$ (at any $h_E > \tilde{h}_E$, it would obtain zero profits). Define $\bar{h}_E = \sup[\text{supp}\Gamma_E]$ and $\bar{h}_I := \sup[\text{supp}\Gamma_I \setminus 1]$. In equilibrium, $U_E^r(\bar{h}_E) = U_I^r(\bar{h}_I)$ must hold.

In any mixed-strategy equilibrium in which the entrant is joined by a positive measure of users, the incumbent must set $h_I = 1$ with positive probability. Suppose, for a contradiction, that the incumbent plays $h_I = 1$ with probability zero. When setting $h_E = \bar{h}_E$, the entrant would only be joined by any user if the incumbent sets $h_I = \bar{h}_I$.[25] Thus, $\Gamma_I$ must have an atom at $\bar{h}_I$ (if it does not, the entrant obtains zero profits when setting $h_E = \bar{h}_E$, a contradiction). But since the entrant never sets a harmful content share above $\bar{h}_E$ and $\Gamma_E$ cannot have an atom at $\bar{h}_E$ (else, both platforms would prefer to set a harmful content share

[25]This is because the entrant sets $h_I = 1$ with probability zero by specification.

slightly below $\bar{h}_p$ rather than $\bar{h}_p$), the incumbent would obtain zero profits when setting $\bar{h}_I$, a contradiction. Thus, the incumbent must play $h_I = 1$. But then, a mixed-strategy equilibrium in which the entrant is joined by a positive measure of users cannot exist under our condition because the incumbent would strictly prefer to set $h_I = \check{h}_I$ instead of $h_I = 1$.

In any mixed-strategy equilibrium, all users must hence join the incumbent, and $h_I^* = \check{h}_I$ must hold. Thus, the equilibrium we have found is unique. ■

# References


D. Acemoglu, D. Huttenlocher, A. Ozdaglar, and J. Siderius. Online business models, digital ads, and user welfare. Technical report, National Bureau of Economic Research, 2024.

H. Allcott, L. Braghieri, S. Eichmeyer, and M. Gentzkow. The welfare effects of social media. American Economic Review, 110(3):629–76, 2020.

H. Allcott, M. Gentzkow, and L. Song. Digital addiction. American Economic Review, 112 (7):2424–2463, 2022.

S. P. Anderson and A. De Palma. Competition for attention in the information (overload) age. The RAND Journal of Economics, 43(1):1–25, 2012.

S. P. Anderson and M. Peitz. Ad clutter, time use, and media diversity. American Economic Journal: Microeconomics, 15(2):227–270, 2023.

G. Aridor, R. Jiménez-Durán, R. Levy, and L. Song. The economics of social media. 2024.

M. Armstrong. Competition in two-sided markets. The RAND journal of economics, 37(3): 668–691, 2006.

G. S. Becker and K. M. Murphy. A simple theory of advertising as a good or bad. The Quarterly Journal of Economics, 108(4):941–964, 1993.

G. Beknazar-Yuzbashev, R. Jiménez-Durán, and M. Stalinski. A model of harmful yet engaging content on social media. In AEA Papers and Proceedings, volume 114, pages 678–683. American Economic Association 2014 Broadway, Suite 305, Nashville, TN 37203, 2024.

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

H. K. Bhargava. If it's enraging, it is engaging: Infinite scrolling in information platforms. 2023.

P. Bordalo, N. Gennaioli, and A. Shleifer. Competition for attention. The Review of Economic Studies, 83(2):481–513, 2016.

M. Bourreau and J. Krämer. Horizontal and vertical interoperability in the dma. available at https://cerre.eu/wp-content/uploads/2023/12/ISSUE-PAPER-CERRE-DEC23DMA-Horiz 2023.

L. Braghieri, R. Levy, and A. Makarin. Social media and mental health. American Economic Review, 112(11):3660–3693, 2022.

E. Brynjolfsson, A. Collis, A. Liaqat, D. Kutzman, H. Garro, D. Deisenroth, and N. Wernerfelt. The consumer welfare effects of online ads: Evidence from a 9-year experiment. Technical report, National Bureau of Economic Research, 2024.

L. Bursztyn, B. Handel, R. Jiménez-Durán, and C. Roth. When product markets become collective traps: the case of social media. working paper, available at https://ssrn.com/abstract=4596071, 2023.

J. Crémer, G. S. Crawford, D. Dinielli, A. Fletcher, P. Heidhues, M. Schnitzer, and F. M. S. Morton. Fairness and contestability in the digital markets act. Yale J. on Reg., 40:973, 2023.

M. Dhakar and J. Yan. Interoperability & privacy: A case of messaging apps. 2024.

DW, 2025. URL https://www.dw.com/en/far-right-afd-appears-as-strongest-german-party-on-t a-69264717.

M. Ekmekci, A. White, and L. Wu. Platform competition and interoperability: The net fee model. Management Science, 2025.

Guardian, 2021. URL https://www.theguardian.com/technology/2021/oct/22/twitter-admits-bias-in-algorithm-for-rightwing-politicians-and-news-outlets.

A. M. Guess, N. Malhotra, J. Pan, P. Barberá, H. Allcott, T. Brown, A. Crespo-Tenorio, D. Dimmery, D. Freelon, M. Gentzkow, et al. How do social media feed algorithms affect attitudes and behavior in an election campaign? Science, 381(6656):398–404, 2023.

R. Hoong. Self control and smartphone use: An experimental study of soft commitment devices. European Economic Review, 140:103924, 2021.

J. Horwitz et al. The facebook files. The Wall Street Journal, available online at: https://www.wsj.com/articles/the-facebook-files-11631713039, 2021.

S. Ichihashi and B.-C. Kim. Addictive platforms. Management Science, 69(2):1127–1145, 2023.

B. Jullien, A. Pavan, and M. Rysman. Two-sided markets, pricing, and network effects. In Handbook of industrial organization, volume 4, pages 485–592. Elsevier, 2021.

M. Kades and F. Scott Morton. Interoperability as a competition remedy for digital networks. Washington Center for Equitable Growth Working Paper Series, 2020.

K. Kamath, A. Sharma, D. Wang, and Z. Yin. Realgraph: User interaction prediction at twitter. In user engagement optimization workshop@ KDD, number ii, 2014.

R. Mosquera, M. Odunowo, T. McNamara, X. Guo, and R. Petrie. The economic effects of facebook. Experimental Economics, 23:575–602, 2020.

A. Prat and T. Valletti. Attention oligopoly. American Economic Journal: Microeconomics, 14(3):530–557, 2022.

J.-C. Rochet and J. Tirole. Platform competition in two-sided markets. Journal of the european economic association, 1(4):990–1029, 2003.

J. N. Rosenquist, F. M. S. Morton, and S. N. Weinstein. Addictive technology and its implications for antitrust enforcement. NCL Rev., 100:431, 2021.

H. E. Sadagheyani and F. Tatari. Investigating the role of social media on mental health. Mental health and social inclusion, 25(1):41–51, 2021.

F. M. Scott Morton and D. C. Dinielli. Roadmap for an antitrust case against facebook. Stan. JL Bus. & Fin., 27:268, 2022.

Statista, 2025. URL https://www.statista.com/statistics/1369032/mental-health-social-media\/-effect-us-users/.

T.-H. Teh, C. Liu, J. Wright, and J. Zhou. Multihoming and oligopolistic platform competition. American Economic Journal: Microeconomics, 15(4):68–113, 2023.

A. L. Wickelgren and D. Gilo. The exclusionary eects of addictive platforms. U of Texas Law, Legal Studies Research Paper (forthcoming), 2024.

# Contestability and the Optimal Regulation of Social Platform Markets

*Online Appendix*



# B Omitted results

## B.1 Auxiliary lemmata characterizing mixed-strategy equilibria

To begin, define $\bar{h}_p := \sup[\text{supp}\Gamma_p \setminus 1]$ and $\underline{h}_p := \inf[\text{supp}\Gamma_p \setminus 1]$ for both $p \in \{E, I\}$.

**Lemma 3.** *In any mixed-strategy equilibrium, there must exist a $h_I \leq \tilde{h}_I$ such that $h_I \in \text{supp}\Gamma_I$. Further, an equilibrium in which there exists no $h_E \leq \tilde{h}_E$ such that $h_E \in \text{supp}\Gamma_E$ can only exist if $e_I^*(\tilde{h}_I) = (1-\rho)e_I^*(1)$ and $U_I^n(\tilde{h}_I) \leq U_E^n(1)$.*

**Proof of Lemma 3:** Suppose, for a contradiction, that there exists an equilibrium in which there is no $h_I \leq \tilde{h}_I$ such that $h_I \in \text{supp}\Gamma_I$. The incumbent would never optimally set a harmful content share $h_I' \in (\tilde{h}_I, 1)$, because it would not be joined by rational users when setting such harmful content shares. Thus, it is superior to set $h_I = 1$ rather than any $h_I' \in (\tilde{h}_I, 1)$ because the incumbent can obtain weakly higher demand and larger engagement by setting $h_I = 1$. Hence, the incumbent must play a pure strategy and set $h_I = 1$ with probability 1.

Since the incumbent plays $h_I = 1$ with probability 1, the entrant would not find it optimal to mix. Note firstly that the entrant would never optimally set the harmful content share $h_E = 1$ (then, it obtains zero demand). The incumbent would also never set $h_E < \tilde{h}_E$, because all rational users join the incumbent for any $h_E \leq \tilde{h}_E$, so the entrant would prefer to set $\tilde{h}_E$ rather than any $h_E < \tilde{h}_E$. Hence, the entrant must play $\tilde{h}_E$ with probability one. This means we are not in a mixed-strategy equilibrium, a contradiction.

Suppose that there exists an equilibrium in which there is no $h_E \leq \tilde{h}_E$ such that $h_E \in \text{supp}\Gamma_E$. By analogous arguments to those made above, the entrant must thus play a pure strategy and set $h_E = 1$ with probability 1.

Thus, there are two harmful content shares which can be optimal for the incumbent: $h_I = \tilde{h}_I$ or $h_I = 1$. If the incumbent mixes, it must be indifferent between these two harmful content shares.

This implies that $U_I^n(\tilde{h}_I) \leq U_E^n(1)$ must hold. Suppose, for a contradiction, that $U_I^n(\tilde{h}_I) > U_E^n(1)$ holds. Then, the entrant would never be joined by any users. But then, it would prefer to set $h_E$ in an open interval above 0 to attract rational users (since this grants the entrant strictly positive profits), a contradiction.

Given that $U_I^n(\tilde{h}_I) \leq U_E^n(1)$ must hold, the incumbent is only joined by rational users when setting $h_I = \tilde{h}_I$. Thus, $\rho e^*(\tilde{h}_I) = (1-\rho)e^*(1)$ must hold. For a given combination of technology parameters, there exists a unique $\rho$ such that this equation is satisfied. Hence, the equilibrium only exists for a parameter space with measure zero. ■

**Lemma 4.** *Consider any MSE in which there exists a* $h'_j \in \text{supp}[\Gamma_j \setminus 1]$ *for both* $j \in \{E, I\}$. *Then,* $U_E^r(\bar{h}_E) = U_I^r(\bar{h}_I)$ *must hold.*

**Proof of Lemma 4:** Suppose, for a contradiction, that there exists such a MSE in which $U_j^r(\bar{h}_j) < U_{-j}^r(\bar{h}_{-j})$ holds, where $j \in \{E, I\}$.

This implies that platform $j$ will, when setting the harmful content share of $h_j = \bar{h}_j$, only be joined by rational users if platform $-j$ sets the harmful content share $h_{-j} = 1$ (this can happen with probability zero). To see why this holds true, note that $U_{-j}^r(h_{-j}) > U_{-j}^r(\bar{h}_{-j})$ holds for any $h_{-j} < \bar{h}_{-j}$.

By implication, $\bar{h}_j = \tilde{h}_j$ must hold: If $\bar{h}_j < \tilde{h}_j$, then platform $j$ would strictly prefer to set $h_j = \tilde{h}_j$ instead of $\bar{h}_j$ because this would weakly increase demand and increase the engagement of platform $j$'s users.

Define $h'_j < \bar{h}_j$ such that $U_j^r(h'_j) = U_{-j}^r(\bar{h}_{-j})$. For any $h_j \in (h'_j, \bar{h}_j)$, the demand which the platform $j$ receives is weakly increasing. Thus, the platform $j$ will never set such a harmful content share , i.e. $h_j \notin \text{supp}\Gamma_j$ holds for all $h_j \in (h'_j, \bar{h}_j)$. This implies that the platform $j$ never sets a harmful content share such that $U_j^r(h_j) \in \big(U_j^r(\bar{h}_j), U_{-j}^r(\bar{h}_{-j})\big)$.

Now consider platform $-j$. Suppose that platform $-j$ deviates by setting a harmful content share slightly above $\bar{h}_{-j}$, i.e. a harmful content share $h_{-j}$ such that $U_{-j}^r(h_{-j}) \in \big(U_j^r(\bar{h}_j), U_{-j}^r(\bar{h}_{-j})\big)$. Then, the demand which platform $-j$ receives from rational users is unchanged, but engagement rises. Thus, the deviation is profitable, a contradiction. ■

**Lemma 5.** *Consider any MSE in which there exists a* $h_j' \in \text{supp}\ \Gamma_j \setminus 1$. *Then,* $1 \in \Gamma_{-j}$ *must have an atom at* $\bar{h}_{-j}$ *or an atom at* 1.

**Proof of Lemma 5 :** Consider any MSE in which there exists a $h_j' \in \text{supp}\Gamma_j \setminus 1$ and suppose, for a contradiction, that $\Gamma_{-j}$ does not have an atom at $\bar{h}_{-j}$ or 1.

Note that there cannot exist a $h_{-j} \in (\tilde{h}_{-j}, 1)$ with $h_{-j} \in \text{supp}\Gamma_{-j}$, because platform $-j$ would obtain strictly higher profits by setting $h_{-j} = 1$ than any such harmful content share.

Now consider platform $j$. When setting $\bar{h}_j$, platform $j$ is thus not joined by any rational users (since its rival always sets a harmful content share at which rational users obtain higher utility, given that $U_j(\bar{h}_j) = U_{-j}(\bar{h}_{-j})$ must hold by Lemma 4). Because it is only joined by naive users when $h_j = \bar{h}_j$, it follows that $\lim_{h_j \to \bar{h}_j} \Pi_j(h_j) < \Pi_j(1)$. Thus, such an equilibrium cannot exist, because platform $j$'s mixing indifference condition cannot be satisfied. ■

**Lemma 6.** *Consider any MSE in which there exists a* $h_E \in \text{supp}\Gamma_E \setminus 1$ *and a* $h_I \in \text{supp}\Gamma_I \setminus 1$. *Then,* $\bar{h}_E = \tilde{h}_E$ *and* $\bar{h}_I = \tilde{h}_I$ *must hold. For both* $j \in \{E, I\}$, *the distribution* $\Gamma_j$ *must be atomless and gapless on* $\left[\inf \text{supp}\Gamma_j, \tilde{h}_j\right)$.

**Proof of Lemma 6 :**

(i) The equality $\bar{h}_j = \tilde{h}_j$ must hold for both $j$.

Suppose, for a contradiction, that $\bar{h}_j < \tilde{h}_j$ holds for some platform $j \in \{E, I\}$. When setting $\bar{h}_j$, the platform is only joined by rational users if its rival sets the harmful content share 1 or $\bar{h}_{-j}$. Define $\lambda_{-j}$ as the probability that platform $-j$ sets $h_{-j} = \bar{h}_{-j}$ or $h_{-j} = 1$.

Suppose $\Gamma_{-j}$ does not have an atom at $\bar{h}_{-j}$. When setting $h_j = \bar{h}_j < \tilde{h}_j$, platform $j$ would only be joined by rational users if its rival sets $h_{-j} = 1$ (this may happen with probability zero). Thus, the demand platform $j$ would obtain when setting $\tilde{h}_j$ is weakly higher than the demand it obtains when setting $\bar{h}_j$, which means that the platform's mixing indifference condition cannot be satisfied, a contradiction.

Suppose $\Gamma_{-j}$ has an atom at $\bar{h}_{-j}$. Then, $\Gamma_j$ cannot have an atom at $\bar{h}_j$. By analogous arguments, it follows that $\bar{h}_{-j} = \tilde{h}_{-j}$ must hold. By the results of Lemma 5, it follows that $\bar{h}_j = \tilde{h}_j$ must hold, since $U_j^r(\bar{h}_j) = U_{-j}^r(\bar{h}_{-j})$ must hold in equilibrium. This is a contradiction.

(ii) The distributions must be atomless.

Suppose, for a contradiction, that the distribution $\Gamma_E$ has an atom at $h'_E \in [\underline{h}_E, \tilde{h}_E)$. There exists a $h'_I > h'_E$ such that rational users are indifferent between either platform if the incumbent sets $h'_I$ and the entrant sets $h'_E$. At this combination of harmful content shares, naive users strictly prefer to join the incumbent. Then, there exists an interval $[h'_I, h'_I + \epsilon)$ such that the incumbent would strictly prefer to set a $h_I$ slightly below $h'_I$ rather than any $h_I$ in this interval, since this triggers an upward jump in demand. Hence, the incumbent will not offer any $h_I \in [h'_I, h'_I + \epsilon)$. But this means that it is not optimal for the entrant to set $h_E$, given that it could raise its harmful content share slightly without reducing its demand from rational users (given that $h_E < \tilde{h}_E$, as specified). This is a contradiction.

Suppose, for a contradiction, that the distribution $\Gamma_I$ has an atom at $h'_I \in [\underline{h}_I, \tilde{h}_I)$. There exists a $h'_E < h'_I$ such that rational users are indifferent between joining either platform if the entrant sets $h'_E$ and the incumbent sets $h'_I$. At this combination of harmful content shares, naive users strictly prefer to join the incumbent. Thus, there exists an interval $[h'_E, h'_E + \epsilon]$ such that, for any $h_E \in [h'_E, h'_E + \epsilon]$, $h_E \notin \text{supp}\Gamma_E$ must hold. This is because the entrant would strictly prefer to set a $h_E$ just under $h'_E$ rather than any $h_E$ in this interval. But this implies that it is not optimal for the incumbent to set $h'_I$, since it could raise its harmful content share slightly without reducing demand from rational users, a contradiction.

(iii) The distributions must be gapless.

Suppose, for a contradiction, that there exists a platform $j$ for which the distribution $\Gamma_j$ has a gap on $[h^1_j, h^2_j]$, i.e. for which $F_j(h^1_j) = F_j(h^2_j)$ holds. Define $h^1_{-j}$ and $h^2_{-j}$ such that $U^r_j(h^1_j) = U^r_{-j}(h^1_{-j})$ and $U^r_j(h^2_j) = U^r_{-j}(h^2_{-j})$. There cannot exist a $h'_{-j} \in (h^1_{-j}, h^2_{-j}]$ in the support of $\Gamma_{-j}$. This implies that $F_{-j}(h^1_{-j}) = \lim_{h_{-j} \to h^2_{-j}} F_{-j}(h_{-j})$ must hold. But this yields a contradiction. Platform $j$ would then prefer to deviate by setting $h^2_j$ instead of a $h_j$ including or slightly below $h^1_j$. ■

**Lemma 7.** *Consider any MSE in which there exists a $h_E \in \text{supp}\Gamma_E \setminus 1$ and a $h_I \in \text{supp}\Gamma_I \setminus 1$. In equilibrium, $U^r_E(\underline{h}^r_E) = U^r_I(\underline{h}^r_I)$ must hold.*

**Proof of Lemma 7 :** Suppose, for a contradiction, that $U^r_j(\underline{h}_j) < U^r_{-j}(\underline{h}_{-j})$ holds for some $j$. When platform $-j$ sets $\underline{h}_{-j}$, all rational users strictly prefer to join platform $-j$. Thus, this platform would prefer to set a $h_{-j}$ slightly above $\underline{h}_{-j}$ rather than $\underline{h}_{-j}$, because this leaves demand from rational users unchanged, weakly increases demand from naive users, and boosts engagement. This is a contradiction. ■

**Lemma 8.** *If $U_I^n(\tilde{h}_I) > U_E^n(1)$, then the following two properties must be satisfied in a mixed-strategy equilibrium:*

- *There exists a $h_E \in \text{supp}\Gamma_E \setminus 1$.*
- $1 \notin \text{supp}\Gamma_E$.

*Thus, any mixed-strategy equilibrium must satisfy the properties laid out in Lemmas 6 - 7. Moreover, $1 \in \text{supp}\Gamma_I$ and $\tilde{h}_E \in \text{supp}\Gamma_E$ must hold.*

**Proof of Lemma 8:**

(i) If $U_I^n(\tilde{h}_I) > U_E^n(1)$, there exists a $h_E \in \text{supp}\Gamma_E \setminus 1$ in any mixed-strategy equilibrium

Suppose, for a contradiction, that there exists a mixed-strategy equilibrium in which there exists no $h_E \in \text{supp}\Gamma_E \setminus 1$. Then, the entrant plays $h_E = 1$ with probability 1. But then, only one of two harmful content shares can be optimal for the incumbent: $h_I = 1$ or $h_I = \tilde{h}_I$. But if the incumbent sets either of these harmful content shares, all naive users join the entrant. Thus, the entrant would obtain zero demand in equilibrium, a contradiction.

(ii) If $U_I^n(\tilde{h}_I) > U_E^n(1)$, then $1 \notin \text{supp}\Gamma_E$ must hold.

Suppose, for a contradiction, that there exists a mixed-strategy equilibrium in which $1 \in \text{supp}\Gamma_E$. By Lemma 6 and previous arguments, $\tilde{h}_I \in \Gamma_I$ must hold. Since $1 \in \text{supp}\Gamma_E$, $\Gamma_E$ must have an atom at $h_E = 1$. When setting the harmful content share $\tilde{h}_I$, the incumbent is joined by naive users even if the entrant sets the harmful content share of 1.

Define $h_I'$ such that $U_I^n(h_I') = U_E^n(1)$. Suppose $\underline{h}_I < h_I'$. Then, the distribution $\Gamma_I$ must have a gap, because the profits which the incumbent obtains jump up at $h_I'$ (given that the incumbent is joined by naive users when the entrant sets $h_E = 1$ if and only if $h_I \geq h_I'$). But the distribution cannot have a gap (Lemma 5), so we have a contradiction.

Suppose instead that $\underline{h}_I \geq h_I'$. Then, the entrant obtains zero demand when setting $h_E = 1$ (and thus, zero profits): Rational users never join it, and naive users always strictly prefer to join the incumbent since $U_I^n(h_I) > U_I^n(h_I') = U_E^n(1)$ holds for almost all $h_I \in \text{supp}\Gamma_I$ and since the distribution $\Gamma_I$ cannot have an atom at $\underline{h}_I$. This is a contradiction.

(iii) If $U_I^n(\tilde{h}_I) > U_E^n(1)$, any mixed-strategy equilibrium must satisfy the properties laid

out in Lemmas 6 - 7. Moreover, $1 \in \text{supp}\Gamma_I$ and $\tilde{h}_E \in \text{supp}\Gamma_E$ must hold.

The first result holds since there exists a $h_E \in \text{supp}\Gamma_E \setminus 1$ and a $h_I \in \text{supp}\Gamma_I \setminus 1$ by previous arguments. Lemma 5 establishes that $\Gamma_E$ must have an atom at $\tilde{h}_E$, since $1 \notin \Gamma_E$. Thus means that $\Gamma_I$ cannot have an atom at $\tilde{h}_I$, so it must have an atom at 1. ■

## B.2 Overview: Derivation of mixed-strategy equilibria under large competitive advantages

To begin with the characterization of the mixed-strategy equilibrium, recall that the incumbent will never optimally set a harmful content share below $\check{h}_I$ and both platforms $p \in \{E, I\}$ will never optimally choose a harmful content share $h_p \in (\tilde{h}_p, 1)$. Moreover, note that both platforms $p \in \{E, I\}$ must set a harmful content level weakly below $\tilde{h}_p$ in a mixed-strategy equilibrium. If the incumbent chooses $h_I = 1$ with probability 1, the uniquely optimal harmful content share for the entrant is $\tilde{h}_E$. If the entrant chooses the harmful content share $h_E = 1$ with probability 1, the incumbent would always either set $h_I = \tilde{h}_I$ or $h_I = 1$, which makes it suboptimal for the entrant to set $h_E = 1$. Furthermore, both platforms also must set a harmful content share weakly above $p \in \{E, I\}$ must set a harmful content level weakly below $\tilde{h}_p$ with strictly positive probability.[26]

The previous arguments imply that the entrant must play the harmful content share $\tilde{h}_E$ with strictly positive probability if $U_E^n(1) \leq U_I^n(\check{h}_I)$. In turn, this implies that the incumbent must play $h_I = 1$ with strictly positive probability.[27]

It remains to establish what harmful content shares below $\tilde{h}_I$ and $\tilde{h}_E$ can be played in equilibrium. To do this, define $\bar{h}_p := \sup[\text{supp}\Gamma_p \setminus 1]$ for both $p \in \{E, I\}$ and $\underline{h}_p := \inf[\text{supp}\Gamma_p \setminus 1]$ for both $p \in \{E, I\}$. Firstly, note that $U_E^r(\bar{h}_E) = U_I^r(\bar{h}_I)$. To see why this must hold, suppose (for a contradiction) that $U_E^r(\bar{h}_E) < U_I^r(\bar{h}_I)$ holds in equilibrium. Then, the entrant is not joined by rational users if it sets the harmful content share $h_E = \bar{h}_E$. The entrant would thus strictly prefer to set the harmful content share $h_E = 1$ instead of $\bar{h}_E$, a contradiction. An analogous argument establishes that $U_E^r(\bar{h}_E) > U_I^r(\bar{h}_I)$ cannot hold in equilibrium. Secondly, note that $U_E^r(\underline{h}_E) = U_I^r(\underline{h}_I)$ must hold in equilibrium. If

[26] Suppose, for a contradiction that some platform $p$ only sets harmful content shares strictly below $\tilde{h}_p$. Now consider the other platform, namely $j$. When setting the harmful content share $h_j = \sup[\text{supp}\Gamma_j \setminus 1]$, the platform will be joined by naive users with probability zero (since its rival never sets $h_p = 1$)

[27] By previous arguments, the incumbent must either play $h_I = 1$ with positive probability of $h_I = \tilde{h}_I$ with positive probability. If the incumbent plays $h_I = \tilde{h}_I$ with positive probability, the entrant would strictly prefer to set a harmful content share just below $\tilde{h}_E$ rather than $\tilde{h}_E$, a contradiction.

$U_E^r(\underline{h}_E) < U_I^r(\underline{h}_I)$ holds, for example, then the incumbent would strictly prefer to set a harmful content share just above $\underline{h}_I$ rather than the harmful content share $\underline{h}_I$.

Taken together, the previous arguments imply that $\text{supp}\Gamma_I = [\underline{h}_I, \tilde{h}_I) \cup 1$ must hold. The lower bound of the support of $\Gamma_I$ must solve the following mixing indifference condition:

$$(1-\rho)e_I^*(1) = e_I^*(\underline{h}_I) \tag{B.1}$$

The left-hand side are the incumbent's profits if it sets $h_I = 1$. The right-hand side are the incumbent's profits it it sets the harmful content share $\underline{h}_I$, given that all users would then join the incumbent.[28]

Moreover, $\text{supp}\Gamma_E = [\underline{h}_E, \tilde{h}_E] \cup 1$ must hold in equilibrium. The lower bound of the support of $\Gamma_E$ is pinned down by the equation

$$U_E^r(\underline{h}_E) = U_I^r(\underline{h}_I). \tag{B.2}$$

Finally, note that $\Gamma_E$ must have an atom at $\tilde{h}_E$ and that $\Gamma_I$ must have an atom at 1. We define the probability that the incumbent sets the harmful content share 1 as $\lambda_I > 0$ and the probability that the entrant sets the harmful content share $\tilde{h}_E$ as $\lambda_E > 0$. The value $\lambda_E$ is set to make the incumbent exactly indifferent between choosing the harmful content shares $\tilde{h}_I$ and 1.

## B.3 Assumption 1 & underlying parameter restrictions

In the following, we show that all parametric examples we consider in Section 4.3. satisfy Assumption 1 and the condition $U_E^n(1) \leq U_I^n(0)$ we impose at the beginning of this section. Recall that we have set $\gamma = 0.25$.

The utility $U_p^n(h_p)$ satisfies:

$$U_p^n(h_p) = \frac{\big((\eta_p - \theta_p)h_p + \theta_p\big)^2}{4\gamma} + 1 - h_p \tag{B.3}$$

(i) The condition $U_E^n(1) \leq U_I^n(0)$ is satisfied for all parameter combinations under consideration.

[28] Suppose the incumbent sets $\underline{h}_I$. Naive users always strictly prefer to join the incumbent if it sets any $h_I \geq \check{h}_I$, given that $U_E^n(1) \leq U_I^n(\check{h}_I)$. Rational users prefer to join the incumbent because $U_E^n(\underline{h}_E) = U_I^n(\underline{h}_I)$ and the probability that the entrant plays a harmful content share strictly above $\underline{h}_E$ is 1.

This condition is implied by the stronger condition $U_E^n(1) \leq U_I^n(0)$. Note that:

$$U_I^n(0) = \frac{(\theta_I)^2}{4\gamma} + 1 = (\theta_I)^2 + 1 \tag{B.4}$$

Note further that $U_E^n(1) = (\eta_E)^2$. Thus, the condition $U_E^n(1) \leq U_I^n(0)$ holds if $\theta_I \geq \eta_E$

In our parametric example, this is always satisfied since $\theta_I = 3$ and $\eta_E \in (1.5, 2.5]$.

(ii) For all parameter combinations we consider, $U_E^n(h_E)$ and $U_I^n(h_I)$ are strictly increasing.

We begin by considering $U_E^n(h_E)$. Note that this function is strictly convex and that

$$\frac{\partial U_E^n(h_E)}{\partial h_E} = \frac{(\eta_E - \theta_E)\big((\eta_E - \theta_E)h_E + \theta_E\big)}{2\gamma} - 1. \tag{B.5}$$

If the derivative of $U_E^n(h_E)$ at $h_E = 0$ is positive, the derivative is generally positive (because $U_E^n(h_E)$ is convex). The derivative of $U_E^n(h_E)$ at $h_E = 0$ is positive if and only if

$$\left.\frac{\partial U_E^n(h_E)}{\partial h_E}\right|_{h_E=0} = \frac{(\eta_E - \theta_E)\big(\theta_E\big)}{2\gamma} - 1 \geq 0 \iff (\eta_E - \theta_E)(\theta_E) \geq 2\gamma \tag{B.6}$$

Since we set $\gamma = 0.25$, this condition is satisfied if $(\eta_E - \theta_E)\theta_E \geq 0.5$. In our examples, this is satisfied because $(2 - 0.5)(0.5) = 0.75$ and because $(2 - 1.5)(1.5) = 0.75$.

Now consider $U_I^n(h_I)$ and recall that we have set $\theta_I = 3$ and $\eta_I = 4$. the derivative of $U_I^n(h_I)$ is generally positive because:

$$\left.\frac{\partial U_I^n(h_I)}{\partial h_I}\right|_{h_I=0} = \frac{(\eta_I - \theta_I)\big(\theta_I\big)}{2\gamma} - 1 \geq 0 \iff 3 \geq 0.5 \tag{B.7}$$

(iii) For all parameter combinations we consider,$U_I^r(h_I)$ and $U_E^r(h_E)$ are strictly decreasing. Moreover, $U_p^r(0) > U_p(1)$ hold for both $p \in \{E, I\}$.

Intuitively, both conditions are satisfied by setting $\delta$ (the multiplier governing to the utility costs of harmful content) large enough.

Note that

$$U_p^r(h_p) = \frac{\big((\eta_p - \theta_p)h_p + \theta_p\big)^2}{4\gamma} + 1 - (1+\delta)h_p. \tag{B.8}$$

The derivative of this w.r.t $h_p$ is negative at $h_p = 1$ (and thus globally negative) if:

$$\left.\frac{\partial U_p^r(h_p)}{\partial h_I}\right|_{h_p=1} = \frac{(\eta_p - \theta_p)\big(\eta_p\big)}{2\gamma} - (1+\delta) < 0 \tag{B.9}$$

For the entrant, we have $\eta_E \in [2, 2.5]$ and $\theta_E \in \{0.5, 1.5\}$. If $\theta_E = 0.5$, we have

$$\left.\frac{\partial U_E^r(h_p)}{\partial h_E}\right|_{h_E=1} \leq \frac{(2.5-0.5)(2.5)}{2(0.25)} - (1+\delta) < 0 \iff 9 < \delta$$

If $\theta_E = 1.5$, we have

$$\left.\frac{\partial U_E^r(h_p)}{\partial h_E}\right|_{h_E=1} \leq \frac{(2.5-1.5)(2.5)}{2(0.25)} - (1+\delta) < 0 \iff 4 < \delta$$

Furthermore, we have $U_E^r(1) = \frac{(\eta_E)^2}{4\gamma} - \delta \leq \frac{(2.5)^2}{0.5} - \delta = 12.5 - \delta$. Thus, setting $\delta > 12.5$ guarantees that this utility is negative.

For the incumbent, we have $\theta_I = 3$ and $\eta_I = 4$. This implies that

$$\left.\frac{\partial U_I^r(h_I)}{\partial h_I}\right|_{h_I=1} = \frac{(\eta_I - \theta_I)(\eta_I)}{2(0.25)} - (1+\delta) < 0 \iff \frac{4}{2(0.25)} - (1+\delta) < 0 \iff 8 < \delta$$

Finally, we have $U_I^r(1) = \frac{(\eta_E)^2}{4\gamma} - \delta \leq \frac{(4)^2}{0.5} - \delta = 16 - \delta$. Thus, setting $\delta > 16$ guarantees that this utility is negative.

Thus, we set $\delta = 20$ to satisfy the desired properties jointly.