Addictive Platform Design: Competition, Awareness, and Regulation*

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

April 17, 2026

Abstract

We build a model of asymmetric competition between social media platforms with behavioral users. Platforms have incentives to amplify harmful content because this increases user engagement, and behavioral users fail to internalize the adverse effects of consuming harmful content. We show that the user-optimal outcome never emerges under asymmetric competition, and that user migration away from a dominant platform can harm users by inducing platforms to display more harmful content. Reducing the share of behavioral users induces both platforms to amplify less harmful content, but may promote market concentration.

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 Panunzi, Anja Prummer, David Ronayne, Nicolas Schutz, Sebastian Schweighofer-Kodritsch, Fiona Scott-Morton, Roland Strausz, Tomasz Sulka, Fernando Vega-Redondo, Ernst-Ludwig von Thadden, Jonas von Wangenheim, and seminar audiences at Berlin, 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

In February 2026, the European Commission has preliminarily found TikTok in breach of the Digital Services Act for its addictive design (European Commission, 2026a). It was noted that features such as infinite scroll, autoplay, and push recommendations could harm the well-being of its users, including minors and vulnerable adults, by fostering compulsive behavior and reducing self-control.[1] Similar legal investigations of dominant platforms such as Shein, Meta, and YouTube are being conducted in the EU and the US.[2] In general, the adverse effects of social media on users' well-being are well-documented (Allcott et al., 2020; Braghieri et al., 2022). In addition, many people utilize social media even though they report that this reduces their well-being (Sagioglou and Greitemeyer, 2014; Hoong, 2021). Social media markets are thus characterized by three key primitives: dominant platforms enjoy competitive advantages, platforms have incentives to amplify content that makes users spend more time on their platform even if this harms users (since this boosts the platforms' revenue), and many users do not internalize the adverse effects of social media.[3] How these primitives interact under platform competition, and what this implies for the design of regulation, is the question this paper studies.

We build a theoretical model of asymmetric competition between a dominant platform (the incumbent) and a rival (the entrant). Platforms simultaneously choose how much to amplify harmful content in their algorithmic feeds. We define harmful content as content that reduces a user's well-being but boosts her engagement, i.e., makes her spend more time on the platform. Based on the extent to which platforms amplify harmful content, users decide which platform to join. The incumbent has a competitive advantage, i.e., any user obtains higher utility by joining the incumbent, ceteris paribus. Importantly, this competitive advantage can be arbitrarily small. Users differ in the degree to which they internalize the adverse effects of harmful content: a share of users are *rational* and account for these effects when choosing which platform to join, while the remaining users are *naive* and do not. We define user welfare as users' expected utility (which accounts for the effects of harmful content). Our specification of user naivety can be viewed as a reduced-form representation

[1] The legal basis of these investigations in the European Union is the Digital Services Act. For example, preamble (83) identifies systemic risks "relating to the design... of very large online platforms ... with an actual or foreseeable negative effect on the protection of public health. Such risks may also stem from ... online interface design that may stimulate behavioural addictions of recipients of the service."

[2] In February 2026, the European Commission opened formal proceedings against Shein, focusing on risks linked to engagement-based design (European Commission, 2026b). The first bellwether trial in California's coordinated proceeding on adolescent social-media addiction (JCCP No. 5225) began in January 2026.

[3] When users spend more time on a platform, it can sell more ad impressions. Since ad revenue is central for platforms (for example, 96.69% of Meta's fourth-quarter revenue in 2024 came from advertising), they have incentives to design engagement-maximizing algorithms (Scott Morton and Dinielli, 2022).

of present bias, which has also been documented among social media users (Hoong, 2019).

Our analysis yields several policy-relevant insights: The user-optimal outcome—in which users are not exposed to harmful content—never emerges if the incumbent has a competitive advantage, while it may emerge if the incumbent has no competitive advantage and there are enough rational users. This suggests that competition policy instruments and initiatives which raise user awareness regarding the adverse effects of social media should be utilized jointly. However, the effects of either policy instrument are non-monotonic. For example, regulation which reduces a dominant platform's competitive advantage can harm users if it induces user migration away from this platform, since this incentivizes it to display more harmful content to its remaining users. Moreover, awareness initiatives can benefit all users by reducing the amount of harmful content platforms display, but may promote the monopolization of these markets if the incumbent's competitive advantage is sufficiently large.

The key trade-off which any platform faces is the following: Amplifying harmful content more increases the engagement of its users, but may induce rational users to leave. A monopoly platform thus optimally displays a small (respectively, large) amount of harmful content if the share of rational users is large (respectively, small). Under asymmetric platform competition, these countervailing incentives give rise to equilibrium regime changes across parameter values, and to equilibria with non-standard structure, including asymmetric mixed-strategy equilibria with gaps and atoms. Such mixed-strategy equilibria arise when the share of rational users is intermediate and neither naive nor rational users constitute a clearly superior source of profit, so that no pure-strategy best response exists. This indeterminacy is economically meaningful: it reflects real-world environments in which small changes in user awareness, platform advantages, or regulatory constraints can tip platforms' behavior from highly harmful to substantially safer content strategies. These equilibria also play a central role in the non-monotonic effects of contestability improvements and welfare.

Our first main result is that user welfare must be strictly larger in an equilibrium in which all users join the incumbent than in any other equilibrium (holding the model's primitives fixed). All users would, for example, obtain weakly negative utility in an equilibrium in which all rational users join the entrant and all naive users join the incumbent: If the market segments like this, the incumbent amplifies harmful content maximally to boost engagement from the naive users who join it. This causes significant harm to naive users and makes it unviable for rational users to join the incumbent, which enables the entrant to extract all surplus from them. By contrast, all users obtain strictly positive utility in an equilibrium where all users join the incumbent—else, the entrant could attract rational users. Importantly, this key result holds independently of the extent of the incumbent's

competitive advantage.

Our second main result is that, if the share of rational users is small, reductions of the incumbent's competitive advantage cannot increase user welfare. This is because the incumbent optimally amplifies harmful content maximally if the share of rational users is small, given that this maximizes the profits it obtains from naive users. Rational users join the entrant but obtain zero surplus because they lack a viable alternative. Improving the entrant's ability to generate utility for its users thus merely enables it to extract more surplus from the rational users it already serves—it does not benefit any user.[4] This establishes that awareness regarding the adverse effects of harmful content must reach a sufficient level before instruments which reduce the incumbent's competitive advantage can improve user welfare.

Our third result formalizes that such awareness initiatives may also face a trade-off: If the share of naive users is small and the incumbent's competitive advantage is sufficiently large, there exists a unique equilibrium in which both platforms display little harmful content, but all users join the incumbent. As the share of rational users grows, the incumbent finds it profitable to moderate its content and attract all users rather than exploiting a shrinking pool of naive users. If its competitive advantage is sufficiently large, it is also visited by naive users even if it moderates its content. Awareness initiatives thus face a tension between reducing harmful content and fostering market concentration, which may have negative long-run effects on innovation and entry.

Content moderation regulation faces similar trade-offs. By constraining the degree to which the incumbent can exploit naive users, this type of regulation strengthens this platform's incentives to moderate content. This directly benefits users, but expands the parameter region in which all users visit the incumbent in equilibrium.

Whether competition benefits or harms users depends on the share of rational users. If this share is small, user welfare is identical in the competitive equilibrium and the monopoly benchmark. When the incumbent has a sufficiently strong competitive advantage, user welfare is strictly higher under competition if the share of rational users is large, but strictly smaller if this share is at an intermediate level. The intuition is as follows: The amount of harmful content the incumbent can display while retaining rational users is smaller under competition than in the monopoly benchmark. If the share of rational users is at an intermediate level, the costs of attracting rational users (in the form of lower engagement) thus outweigh its benefits under competition, but not under monopoly. Under competition, the incumbent thus optimally foregoes most rational users and displays a lot of harmful content, while it moderates its content to attract rational users in the monopoly benchmark. If the

[4] Analogously, reducing the incumbent's technological capability in this regime does not alter the equilibrium structure but directly harms its users by lowering the utility generated by non-harmful content.

share of rational users is large, the incumbent always finds it optimal to attract rational users, so it amplifies harmful content less under competition.

Our results suggest that regulatory measures which reduce the incumbent's competitive advantage, content moderation policies, and awareness initiatives should be designed jointly, as their effects are interdependent. Competition policy instruments that reduce the incumbent's competitive advantage, such as data portability mandates (as codified in the EU GDPR and the DMA) and interoperability requirements[5] cannot benefit users when awareness is limited. Moreover, if such instruments induce user migration away from a dominant platform, the resulting increase in harmful content amplification should be constrained by content moderation provisions such as those in the DSA. Social media warning labels can unfold beneficial effects by increasing the share of users who are rational in our sense, but may foster market dominance unless the incumbent's competitive advantage is small enough.

Although our model is stylized, it speaks to current policy debates on age-based restrictions and the effects of AI. Age-based restrictions such as Australia's recent ban on social media for users under 16 may have an additional positive externality on users not affected by the ban.[6] This is because denying minors access to social media raises the share of rational users among the remaining potential user base, which may incentivize platforms to display less harmful content. Generative AI can amplify the relevance of the mechanisms we uncover by making harmful yet engaging content cheaper to produce and potentially harder for users to identify (Menczer et al., 2023). The effect of AI on dominant platforms' competitive advantages is ambiguous: if data is portable, AI may boost entrants' ability to provide engaging content. If access to data is concentrated, developments in AI may instead reinforce incumbency advantages by raising the value of proprietary data.

Our key results extend under various modifications of our baseline model, e.g., if users can multi-home, under network effects, if users misperceive the extent to which platforms amplify harmful content, or if there is a continuum of types that vary in the extent to which they internalize the adverse effects of harmful content. This is because the key mechanisms we uncover are robust and are also of first order in generalized settings: User migration away from a dominant platform goes along with greater market segmentation, which enables both platforms to amplify harmful content more. If a dominant platform chooses to amplify harmful content less, this induces rational users to join it—thus, reductions in the prevalence of harmful content naturally go along with increased market concentration.

[5]The DMA's interoperability provisions (Article 7) currently apply only to Number-Independent Interpersonal Communication Services (NIICS), not to social media feeds. However, the regulatory logic of reducing incumbents' competitive advantages extends to content-based platforms.

[6]See Australia's Online Safety Amendment (Social Media Minimum Age) Act of 2024. Similar proposals are under consideration in several other jurisdictions.

Related Literature: We contribute to the literature on platform competition by considering a model of competition between content creation platforms with two key features of social media markets, namely platform asymmetries and the presence of users who do not internalize the adverse effects of consuming harmful content. Existing work on platform competition almost entirely abstracts from the presence of behavioral users, and usually considers symmetric platforms. Our key insights, e.g., that user migration away from a dominant platform boosts platforms' incentives to amplify harmful content and that awareness initiatives may foster market dominance, emerge through the interaction of these two features.

There is a large literature on platform competition that studies how platforms set prices and advertising policies in the presence of network effects (e.g., Rochet and Tirole, 2003; Caillaud and Jullien, 2003; Parker and Van Alstyne, 2005; Armstrong, 2006; Anderson and De Palma, 2012; Ambrus et al., 2016; Bordalo et al., 2016; Teh et al., 2023; Anderson and Peitz, 2023; Ekmekci et al., 2025).[7] In contrast to most papers in the literature, we consider platforms that compete via content provision rather than pricing or ad loads. Prat and Valletti (2022) consider a model of competition with symmetric platforms that leverage information about users' preferences to sell targeted advertising to firms. In contrast to our work, Prat and Valletti (2022) abstract from the presence of behavioral users, and focus on downstream firms' competition for user attention.

Anderson and Coate (2005) and Peitz and Valletti (2008) consider media platforms which compete to attract users through content provision and the display of advertisements. When platforms increase ad loads, this boosts their revenue, but may reduce their demand. This trade-off is similar to the countervailing effects of amplifying more harmful content in our model, and may give rise to competitive equilibria in which there is too much advertising from a social welfare perspective.[8] Going beyond this, we study a model with asymmetric platforms and in which some users disregard the adverse effects of harmful content.

Our focus on content provision is motivated by the evidence in Brynjolfsson et al. (2024), who show that advertisements only have limited effects on user welfare. This suggests that the harms of social media operate primarily through the content that is displayed and the time this makes users spend online. This is in line with the assessment that platforms amplify harmful content because this maximizes engagement (e.g., Scott Morton and Dinielli, 2022).

Ichihashi and Kim (2023) and Beknazar-Yuzbashev et al. (2024) study the incentives of symmetric platforms to provide harmful yet engaging content. We build on their work

[7] Hagiu and Jullien (2014) study how competition shapes platforms' incentives to guide consumer search.

[8] Relatedly, Anderson and Peitz (2023) show that entry can increase ad clutter and worsen the user experience on all platforms.

by modeling platform asymmetries and the presence of behavioral users, which are key features of social media markets. Bhargava (2023) studies how the entry of a platform that exogenously displays no harmful content affects an incumbent platform's incentives to amplify addictive content, and Wickelgren and Gilo (2024) show how the provision of addictive content can deter entry.

Boundedly rational users remain largely absent from the theoretical literature on platform competition, as noted 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 conveyed by ads about product quality. In contrast to our work, Acemoglu et al. (2024) only consider symmetric platforms and study a different form of bounded rationality.

A growing empirical literature documents that social media use can reduce user welfare (Mosquera et al., 2020; Allcott et al., 2020; Horwitz et al., 2021; Braghieri et al., 2022).[9] At the same time, users do not appear to fully internalize these adverse effects when deciding which platforms to join and how much time to spend on them (Hoong, 2021; Allcott et al., 2022).[10] Allcott et al. (2022) show that these behavioral patterns can emerge due to habit formation and imperfect self-control. Sagioglou and Greitemeyer (2014) document that users expect to feel better after using Facebook even though this actually makes them feel worse.[11]

Our paper is also broadly related to a literature in behavioral IO which studies markets where consumers neglect shrouded fees or evaluate product quality incorrectly, e.g., Gabaix and Laibson (2006), Heidhues et al. (2016), and Heidhues and Kőszegi (2017). A common theme in this literature is that competition does not necessarily protect consumers from exploitation. Our work differs from the papers in this literature by considering a model of asymmetric competition, and with a different form of user naivety.

Finally, our paper speaks to a growing literature on the regulation of digital platforms. Kades and Scott Morton (2020) argue that interoperability is essential for healthy platform competition. Giovannetti and Siciliani (2023) establish how the interaction of switching costs and network effects can foster market dominance. Bourreau and Krämer (2023) and Dhakar and Yan (2024) show that horizontal interoperability may reduce multihoming and weaken competitive pressure, and may also reduce platform quality. Jeon and Rey (2026) establish that promoting competition or interoperability in app-store environments can raise commissions and reduce innovation. The mechanisms in these papers are different from ours, given that they abstract from the presence of behavioral users.

[9] The potential harmful effects of advertising were already emphasized by Becker and Murphy (1993).

[10] Bursztyn et al. (2025) emphasize that users may experience disutility from not participating on social media, often referred to as fear of missing out.

[11] Algorithmic personalization can further increase engagement at the cost of welfare; see, for example, Guess et al. (2023); Beknazar-Yuzbashev et al. (2025); Risco and Lleonart-Anguix (2024).

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 and the entrant.

Users' actions: Any user must decide which platform to join (we show that our results extend if users can multi-home in Section C.1 of the online appendix). 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 $e_i \in \mathbb{R}$.

Platforms' actions: Every platform chooses the share of harmful (yet engaging) content it displays 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.[12] We abstract from the presence of network effects throughout our main analysis, but show that our insights also extend if there are network effects (see Section C.2 of the online appendix).

A fraction $\rho \in (0, 1)$ 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 simplicity, we assume that the chosen engagement levels of a rational and a naive user on a given platform are 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 $e_p^*(h_p)$, where $e_p^*(h_p) \geq 0$ holds for any $h_p \geq 0$. In Section C.3 of the online appendix, we consider an extension in which there are differences between the engagement levels of rational and naive users on a platform.

A rational user maximizes her utility through the choice of j_i. Specifically, a rational user joins platform l instead of platform k if $V_l(h_l) \geq V_k(h_k)$ and $V_l(h_l) \geq 0$, where $V_p(h_p) := U_p(h_p, e_p^*(h_p))$. Naive users join the platform on which they obtain higher perceived utility, which we refer to as $V_p^n(h_p)$. We impose the following assumptions which are in line with our understanding of harmful (yet engaging) content on social media platforms:

Assumption 1. *The following assumptions hold:*

1. *The incumbent platform has a competitive advantage:* $V_I(h) > V_E(h)$ *and* $V_I^n(h) > V_E^n(h)$ *hold for all* $h \in [0, 1]$.

[12] Our insights naturally extend to settings with negative outside options as in Bursztyn et al. (2025).

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

3. *Naive users neglect the adverse effects of harmful content: For both $p \in \{I, E\}$, $V_p^n(h)$ is continuous and weakly increasing in h.*

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

We define user welfare as users' expected utility. The true utility a naive user attains on platform p is $V_p(h_p)$ because all users' engagement levels on a given platform are the same.

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

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

For every $p \in \{I, E\}$ and every $t \in \{r, n\}$, $\pi_p^t(x)$ is a strictly increasing and continuous function. Our flexible specification of $\pi_p^t(x)$ implies that a platform may 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.

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:

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

We adopt the tie-breaking rule that a naive user joins the incumbent if she obtains the same perceived utility when joining either platform. This tie-breaking rule simplifies the analysis by disciplining the behavior of naive users in mixed-strategy equilibria when their perceived utility is flat. Importantly, however, our results also extend under alternative tie-breaking rules, e.g., if naive users join either platform with equal probability when indifferent.

Interpretation of our model: We now highlight key features of the model and explain how it maps to central features of social media platform markets.

Model of social media platforms: We study platforms on which users consume content selected and ranked by algorithmic feeds. Platforms monetize user attention, so revenues depend on the total engagement generated by their users. This setting maps naturally to competition between platforms such as Instagram and TikTok or 𝕏 and Meta Threads, and also applies to content-sharing platforms such as YouTube (Aridor et al., 2024).

Platform technology and incumbency advantage: A platform's ability to generate utility for its users depends on the quality of its recommendation technology and on complementary assets such as user data and the size of its user base. 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 better access to data and a larger user base.

Advertising: Although we do not model advertising explicitly, our flexible specification of platform revenue captures the role of advertising in reduced form. In ad-funded platforms, more engagement generates larger revenues by increasing the amount of ad impressions that can be sold and by improving targeting opportunities.

Harmful content: In line with the assessment of the European Commission concerning the addictive design of platforms such as TikTok, we define harmful content as content that lowers users' utility while increasing engagement (Assumption 1, points 2 and 4). Examples include content which induces compulsive use (e.g., content that promotes doomscrolling), emotionally charged or polarizing content which increases engagement while creating anxiety and stress (e.g., content on divisive issues), and content that is engaging but may harm well-being through social comparison (e.g., highly curated fitness or beauty content). This interpretation is consistent with related notions of addictive content in Ichihashi and Kim (2023) and harmful yet engaging content in Beknazar-Yuzbashev et al. (2024).

User heterogeneity and naivety: A key assumption of our model is that there are users whose perceived utility of joining a platform is weakly increasing in the harmful content share chosen by this platform (Assumption 1, point 3). This specification of naivety can be interpreted as a reduced-form representation of present bias: The positive effects of social media usage on mood and self-esteem tend to be instantaneous and temporary (Marciano et al., 2022; Dreier et al., 2024). By contrast, the adverse effects of social media (e.g., on mental health or isolation) are only experienced gradually over time. Individuals that exhibit strong degrees of present bias would thus largely ignore the harmful effects of social media usage in their decision-making, and behave just as our naive users.

Alternatively, these behavioral patterns may emerge due to limited attention with respect to habit formation (Allcott et al., 2022), or due to the belief that the harmful effects of social media apply to others but not to oneself—a pattern that is prevalent among underage

social media users in the US (PEW Research, 2025). For the equilibrium analysis, the underlying sources of consumer naivety are immaterial, while these are naturally important for interventions which aim to increase the share of rational users as we define them.

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 on this in Section 6.4.

Monopoly benchmark: Before presenting the equilibrium analysis, we briefly characterize the optimal behavior of a monopolist platform. This illustrates the key trade-off that platforms also face under competition: Each platform seeks to maximize the engagement of users who join it, which incentivizes it to display more harmful content. However, the platform will not obtain any revenue from rational users if it displays too much harmful content, because these users will not join the platform in that case.

Formally, suppose the incumbent platform is a monopolist. Then, it will either optimally choose $h_I = \tilde{h}_I$ or $h_I = 1$, where $\tilde{h}_I$ is the level of harmful content at which a rational user is exactly indifferent between joining the incumbent and no platform at all. This cutoff solves

$$V_I(\tilde{h}_I) = 0. \tag{2}$$

By choosing $h_I = \tilde{h}_I$, the incumbent ensures that it is joined by all users, whereas choosing $h_I = 1$ maximizes the engagement of naive users. It is optimal for the incumbent to set $h_I = 1$ if the share of rational users is small enough, i.e., if and only if

$$(1-\rho)\pi_I^n(e_I^*(1)) \geq \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)). \tag{3}$$

3 Equilibrium Analysis

3.1 General Analysis

In this subsection, we provide the equilibrium analysis of the model we laid out in Section 2. The key take-aways from this analysis are the following: Firstly, user welfare is strictly larger in any equilibrium in which all users visit the incumbent than in any equilibrium in which the entrant is visited by some users (holding users' preferences fixed). This result holds

independently of the extent of the incumbent's competitive advantage. Secondly, reductions of a dominant platform's competitive advantage cannot raise user welfare if the share of naive users is large. Thirdly, reducing the share of naive users to negligible levels can benefit all users by reducing the amount of harmful content platforms display. However, such measures will grant the dominant platform a market share of one if its competitive advantage is large.

We begin by characterizing the user-optimal outcome and establish that this outcome never emerges in equilibrium if the incumbent has a competitive advantage:

Lemma 1 (User-optimal outcome)**.**
User welfare is maximal if all users visit the incumbent and the incumbent chooses $h_I = 0$ with probability 1. This outcome never emerges in equilibrium.

To see why user welfare is maximal if all users join the incumbent and this platform displays no harmful content, note firstly that any user would obtain higher utility by joining the incumbent if both platforms choose the same harmful content share (point 1 of Assumption 1). This feature, together with the fact that the utility a user obtains on a given platform is maximal if this platform displays no harmful content (point 2 of Assumption 1) implies the first result in the Lemma. The second result in the Lemma holds because the incumbent would never optimally set the harmful content share $h_I = 0$. This is because rational users strictly prefer to visit the incumbent if $h_I = 0$, which means that it could increase its engagement without reducing the size of its user base by marginally increasing h_I.

Before moving forward, we define the objects $\tilde{h}_E$ and $\check{h}_I$: A user obtains zero utility if she joins the entrant and this platform displays the harmful content share $\tilde{h}_E$, i.e.,

$$V_E(\tilde{h}_E) = 0. \tag{4}$$

Recall that $\tilde{h}_I$ was defined analogously in equation (2). The value of $\check{h}_I$ is the harmful content share at which rational users are indifferent between joining the incumbent and the entrant if the entrant displays no harmful content and the incumbent chooses the harmful content share $\check{h}_I$. Formally, $\check{h}_I \in (0, 1)$ solves the equation

$$V_E(0) = V_I(\check{h}_I). \tag{5}$$

Note that $\check{h}_I < \tilde{h}_I$ holds. To see why, observe that a user must obtain strictly positive utility by visiting the incumbent if this platform sets $\check{h}_I$ (since $V_E(0) > 0$ holds by assumption), whereas she obtains zero utility if the incumbent sets $\tilde{h}_I$. Given that the utility of visiting any platform is falling in the platform's harmful content share, $\check{h}_I < \tilde{h}_I$ must thus hold.

As is standard, we refer to equilibria in which all players play a pure strategy as pure-strategy equilibria, and to any other equilibrium as a mixed-strategy equilibrium. The share of harmful content the incumbent (respectively, the entrant) chooses in a pure-strategy equilibrium is labeled as h_I^* (respectively, h_E^*). The distribution of the harmful content shares the incumbent (respectively, the entrant) chooses in a mixed-strategy equilibrium is labeled as Γ_I (respectively, Γ_E). We say that an equilibrium is unique if all equilibria that exist are outcome-equivalent, i.e., yield the same profits for each platform and the same expected utility for users. We also note that there is no equilibrium in which all users join the entrant because the incumbent has a competitive advantage (point 1 of Assumption 1).

In the following, we establish that user welfare is strictly larger in any equilibrium in which all users visit the incumbent with probability 1 than in any other equilibrium. To build intuition, we first establish this result for the set of pure-strategy equilibria:

Lemma 2 (Equilibrium effects of user migration: Pure-strategy equilibria)**.**
In a pure-strategy equilibrium in which all users join the incumbent, all users obtain strictly higher utility than in any pure-strategy equilibrium in which some users join the entrant.

The logic which underlies this result is the following: In any equilibrium in which the incumbent is joined by all users, all users must obtain strictly positive utility by joining the incumbent. Otherwise, the entrant would deviate from the equilibrium by choosing a harmful content share at which users would obtain strictly positive utility by joining it because this induces rational users to join the entrant (thereby granting it profits above the profits it obtains in equilibrium).

In any other pure-strategy equilibrium, all naive users join some platform $l \in \{I, E\}$ and all rational users join another platform $p \neq l$. Because platform l is only joined by naive users, this platform optimally displays the maximal share of harmful content (so the naive users who visit it obtain strictly negative utility). By implication, rational users never optimally join platform l. In turn, this implies that rational users must, in equilibrium, obtain zero utility when joining platform p. Otherwise, this platform could obtain strictly higher profits by slightly increasing the share of harmful content it shows.[13] Together, these arguments establish that the utility of all users must be strictly larger in a pure-strategy equilibrium in which all users join the incumbent than in any other pure-strategy equilibrium—all users obtain strictly positive utility in the former, and weakly negative utility in the latter.

Importantly, this result holds independently of the extent of the incumbent's competitive advantage. This is because it emerges due to the platforms' endogenous responses to user

[13]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.

migration: If rational users leave the incumbent because the entrant provides a healthier environment, for example, the incumbent will amplify harmful content more aggressively to extract more revenue from its remaining naive users. In turn, this induces the entrant to also amplify harmful content more, since rational users now lack a viable alternative.

We now generalize the result of Lemma 2 by showing that it also extends if we consider mixed-strategy equilibria.

Proposition 1 (Equilibrium effects of user migration)**.**
In an equilibrium in which all users join the incumbent with probability 1, user welfare must be strictly larger than in any equilibrium in which some users join the entrant.

To understand the result, note firstly that any equilibrium in which all users visit the incumbent with probability 1 and the pure-strategy equilibrium in which all users visit the incumbent must be outcome-equivalent. To see why this holds true, consider an equilibrium in which all users visit the incumbent with probability 1 (which can be in pure or mixed strategies). In such an equilibrium, the incumbent must choose the harmful content share $\check{h}_I$ with probability 1. If the incumbent ever chooses a harmful content share above $\check{h}_I$, the entrant would deviate from the equilibrium by choosing a harmful content share just above 0. This is because the entrant could obtain positive profits through this deviation, whereas it obtains zero profits in the equilibrium under consideration (given that all users visit the incumbent with probability 1). If the incumbent sometimes chooses a harmful content share below $\check{h}_I$, it could increase the amount of harmful content it displays without reducing its user base (which is profitable because this raises engagement). In any equilibrium in which all users visit the incumbent with probability 1, all users thus obtain the utility $V_I(\check{h}_I)$.

In any mixed-strategy equilibrium in which the entrant is joined by some users, all users obtain a utility weakly below $V_I(\check{h}_I)$, and some users obtain a utility strictly below $V_I(\check{h}_I)$. To see this, note that the incumbent must choose a harmful content share strictly above $\check{h}_I$ with positive probability in any mixed-strategy equilibrium in which the entrant is visited by some users: If the incumbent chooses $\check{h}_I$ with probability 1, then the entrant would find it optimal to choose $h_E = 1$.[14] But then, the incumbent would not find it optimal to choose $\check{h}_I$. Any user who joins the incumbent when it displays a harmful content share above $\check{h}_I$ obtains a utility strictly below $V_I(\check{h}_I)$ because the utility a user obtains is decreasing in the share of harmful content she consumes. In addition, the utility a user obtains when she joins the entrant must be weakly below $V_I(\check{h}_I)$, given that $V_E(0) = V_I(\check{h}_I)$ holds by definition.

[14]To see this, consider an equilibrium in which some users join the entrant and the incumbent sets $\check{h}_I$ with probability 1. Then, the measure of rational users who visit the entrant in equilibrium must be zero. As a result, the entrant finds it strictly optimal to choose $h_E = 1$.

Taken together, the previous arguments establish the desired result: In any equilibrium in which all users visit the incumbent with probability 1, user welfare must be larger than in any other pure-strategy equilibrium (by Lemma 2) and larger than in any other mixed-strategy equilibrium (by the arguments in the previous paragraph).

We now provide a more detailed characterization of the possible pure-strategy equilibria that can emerge within our model:

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

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

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*. Recall that there exists no equilibrium in which all users visit the entrant because the incumbent has a competitive advantage.

The results of Proposition 2 then hold by the following arguments: In any equilibrium in which all naive users join a platform $p \in \{E, I\}$ and all rational users join the platform $l \neq p$, $h_p^* = 1$ and $h_l^* = \tilde{h}_l$ must hold. This is because any platform that is visited by all naive users (but no rational users) will display maximal harmful content, so its rival will optimally extract all surplus from the rational users who visit it. Thus, there is a unique candidate for a pure-strategy equilibrium in which all rational users visit the entrant and all naive users visit the incumbent (equilibrium candidate 1) and a unique candidate for a pure-strategy equilibrium in which all rational users visit the incumbent and all naive users visit the entrant (equilibrium candidate 2). Moreover, there is a unique candidate for a pure-strategy equilibrium in which all users join the incumbent (equilibrium candidate 3) by the arguments made under the statement of Proposition 1.

We now characterize the equilibrium outcomes that emerge in two specific situations, namely when (i) the share of rational users is small, and (ii) when the competitive advantage of the incumbent is large and the share of rational users is large. In situation (i), the naivety-focused equilibrium is the unique equilibrium. In situation (ii), the market dominance equilibrium is the unique equilibrium. To understand the following results, recall that ρ is the share of rational users.

Proposition 3 (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 the share of naive users is large enough (i.e., $\rho < \underline{\rho}$ holds), the incumbent finds it optimal to entirely focus on naive users and set a harmful content share of one—by doing so, the incumbent ensures that it is visited by all naive users and obtains maximal revenue from them. Then, the best choice the entrant has is to set $h_E = \tilde{h}_E$ and attract rational users. If the share of rational users is small enough, setting $h_I^* = 1$ and $h_E^* = \tilde{h}_E$ is thus uniquely optimal for the platforms, which means that there exists a unique equilibrium in which $h_I^* = 1$ and $h_E^* = \tilde{h}_E$.

If the share of rational users is below $\underline{\rho}$, platform entry does not affect the utility of users. This is because the incumbent chooses $h_I^* = 1$ in the competitive equilibrium and in the monopoly benchmark if $\rho < \underline{\rho}$, given that it strictly prefers to choose the harmful content share 1 rather than the harmful content share $\tilde{h}_I$ (the most profitable deviation in both situations) if $\rho < \underline{\rho}$. All users thus obtain the same utility in the competitive equilibrium and the monopoly benchmark: Rational users obtain zero utility in the competitive equilibrium and do not visit any platform in the monopoly benchmark (which grants them zero utility). Naive users visit the incumbent and obtain the utility $V_I(1) < 0$ in either situation.

The results of Proposition 3 also suggest that there are profound complementarities between regulation that reduces the incumbent's competitive advantage and initiatives that promote awareness regarding the adverse effects of harmful content: If the share of rational users is small, reducing the incumbent's competitive advantage cannot benefit users. This is because naive users always join the incumbent and consume maximal harmful content in equilibrium, 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 only enables the entrant to retain rational users even when displaying more harmful content. Analogously, reducing the incumbent's ability to generate utility will only reduce the utility of naive users.

Throughout the following analysis, we consider a setting in which the incumbent has a relatively strong competitive advantage. Formally, we assume that $V_E^n(1) < V_I^n(\check{h}_I)$. This guarantees that all naive users visit the incumbent in equilibrium, given that their perceived utility of visiting the entrant must be below $V_E^n(1)$, whereas their perceived utility of visiting the incumbent must be above $V_I^n(\check{h}_I)$ in any equilibrium.[15]

The following proposition characterizes the equilibria that emerge, depending on the

[15] To see why the latter holds, note that the incumbent would never choose a harmful content share below $\check{h}_I$ in equilibrium.

share of rational users (ρ). We define a threshold $\bar{\rho} \in (0,1)$ such that $(1-\bar{\rho})\pi_I^n(e_I^*(1)) = \bar{\rho}\pi_I^r(e_I^*(\check{h}_I)) + (1-\bar{\rho})\pi_I^n(e_I^*(\check{h}_I))$. This means that the incumbent would prefer to set $h_I^* = \check{h}_I$, provided this attracts all users, rather than $h_I^* = 1$ if $\rho > \bar{\rho}$. Further, recall that $\underline{\rho}$ is the value of ρ such that the naivety-focused equilibrium is the unique equilibrium if $\rho < \underline{\rho}$.

Proposition 4 (Equilibrium existence and uniqueness)**.**
Suppose $V_E^n(1) < V_I^n(\check{h}_I)$.

- *If* $\rho \leq \underline{\rho}$*, there exists a unique equilibrium in which* $h_I^* = 1$ *and* $h_E^* = \tilde{h}_E$.
- *If* $\rho \in (\underline{\rho}, \bar{\rho})$*, there exists a unique equilibrium in which platforms play mixed strategies with* $\text{supp}\Gamma_I = [\underline{h}_I, \tilde{h}_I] \cup \{1\}$ *and* $\text{supp}\Gamma_E = [\underline{h}_E, \tilde{h}_E]$*, where* $\underline{h}_I$ *solves* $(1-\rho)\pi_I^n(e_I^*(1)) = \rho\pi_I^r(e_I^*(\underline{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\underline{h}_I))$ *and* $\underline{h}_E$ *solves* $V_E(\underline{h}_E) = V_I(\underline{h}_I)$.
- *If* $\rho > \bar{\rho}$*, there exists a unique equilibrium in which* $h_I^* = \check{h}_I$ *and* $h_E^* = 0$.

If the share of rational users is small, the naivety-focused equilibrium emerges by the previously discussed logic. If the share of rational users is large (and the incumbent's competitive advantage is sufficiently large), the market dominance equilibrium emerges. The underlying intuition is as follows: If the share of rational users is large, the incumbent finds it uniquely optimal to choose a harmful content share that is low enough to ensure it is visited by all rational users (namely, $\check{h}_I$). If the incumbent's competitive advantage is sufficiently large (formally, if $V_E^n(1) < V_I^n(\check{h}_I)$ holds), all naive users then also prefer to visit the incumbent in equilibrium, regardless of the harmful content share chosen by the entrant.

If the share of rational users is intermediate, a mixed-strategy equilibrium emerges. The entrant displays harmful content shares in $[\underline{h}_E, \tilde{h}_E]$ in equilibrium, where $\underline{h}_E$ solves

$$V_E(\underline{h}_E) = V_I(\underline{h}_I), \tag{6}$$

which implies that all rational users visit the entrant if it sets $h_E = \underline{h}_E$. The incumbent displays harmful content shares in the set $[\underline{h}_I, \tilde{h}_I] \cup 1$, where $\underline{h}_I$ solves

$$\rho\pi_I^r(e_I^*(\underline{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\underline{h}_I)) = (1-\rho)\pi_I^n(e_I^*(1)), \tag{7}$$

which ensures that it is indifferent between choosing $\underline{h}_I$ and 1.

For convenience, we now provide a visualization of the supports of Γ_I and Γ_E in the mixed-strategy equilibrium. A circle at a harmful content share indicates that the distribution of harmful content shares has an atom at this point:

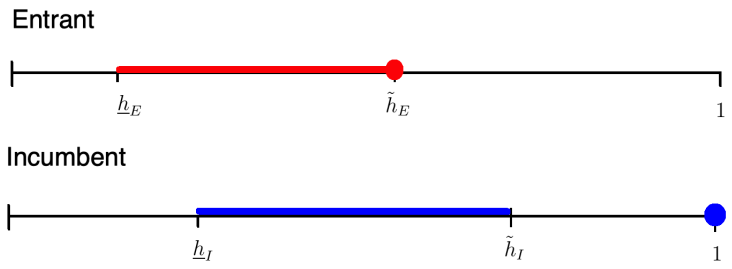

Figure 1: Visualization: Mixed-strategy equilibria

We complete the analysis in this subsection by pinning down the effects of competition on user welfare (under the assumption that the incumbent's competitive advantage is sufficiently strong). Interestingly, the effect is non-monotonic in the share of rational users (ρ).

Proposition 5 (The effects of competition)**.**
If $V_E^n(1) < V_I^n(\check{h}_I)$, the effects of competition on user welfare can be characterized as follows:

- *If $\rho < \underline{\rho}$, user welfare is identical in the monopoly benchmark and the competitive equilibrium.*
- *There exists a $\rho^m > \underline{\rho}$ such that, if $\rho \in (\underline{\rho}, \rho^m)$, user welfare is strictly larger in the monopoly benchmark than in the competitive equilibrium.*
- *If $\rho > \bar{\rho}$, user welfare is strictly lower in the monopoly benchmark than in the competitive equilibrium.*

Thus, competition leaves user welfare unaffected if the share of rational users is small, harms users if the share of rational users is at an intermediate level, and benefits users if sufficiently many users are rational. To understand this result, recall that $\underline{\rho}$ is the threshold level of ρ at which the market transitions from the naivety-focused equilibrium to the mixed-strategy equilibrium, and $\bar{\rho}$ is the threshold level of ρ at which the market transitions from the mixed-strategy equilibrium to the market dominance equilibrium. Note further that $\underline{\rho}$ solves $(1-\underline{\rho})\pi_I^n(e_I^*(1)) = \underline{\rho}\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\underline{\rho})\pi_I^n(e_I^*(\tilde{h}_I))$. This means that the incumbent sets $h_I^* = 1$ if and only if $\rho < \underline{\rho}$, both in the monopoly benchmark and under competition.

When the share of rational users is small, competition has no effect on user welfare. This is because the incumbent always chooses the maximal harmful content share and naive users visit it, while the entrant would set $\tilde{h}_E$ in the competitive equilibrium, so rational users would attain zero utility both in the monopoly benchmark and under competition. When the share of rational users is large, competition benefits users. The key reason is that the incumbent chooses its harmful content share to induce rational users to join it. Under monopoly, the

incumbent can achieve this by setting $\tilde{h}_I$, i.e., displaying a relatively large amount of harmful content at which rational users would attain zero utility. Under competition, this is no longer feasible, so the incumbent chooses a lower harmful content share, which benefits all users.

If the share of rational users is at an intermediate level—specifically, just above $\underline{\rho}$—competition harms users. The underlying intuition is the following: If the share of rational users is just above $\underline{\rho}$, the incumbent would optimally choose the relatively low harmful content share $\tilde{h}_I$ in the monopoly benchmark. This is optimal because this induces all users to visit it, which maximizes its profits because $\rho > \underline{\rho}$. Under competition, however, rational users would not visit the incumbent in equilibrium if it sets $\tilde{h}_I$, given that the entrant would poach rational users in this case. This implies that the incumbent prefers to display relatively large harmful content shares under competition, because it requires very low harmful content shares to attract rational users (which yield little engagement for the incumbent). Thus, the incumbent displays more harmful content under competition than in the monopoly benchmark, which harms users.

3.2 A parametric example & comparative statics

In this section, we complement the insights of the previous section by considering a particular parametric example of the general preference framework we laid out in Section 2. Specifically, we suppose that the true utility a user attains when joining a platform $p \in \{E, I\}$ is

$$U_p(h_p, e_i) = \big(\eta_p h_p + \theta_p(1 - h_p)\big)e_i + (1 - h_p) - \delta h_p - \gamma(e_i)^2, \tag{8}$$

where h_p is the platform's chosen harmful content share and $e_i \in \mathbb{R}$ is the user's engagement level. The platform displays a share $1 - h_p$ of non-harmful content. The parameter $\eta_p \geq 0$ (respectively, $\theta_p \geq 0$) governs the extent to which the consumption of harmful content (respectively, non-harmful content) generates utility via its interaction with engagement through instantaneous gratification. The parameters $\delta > 0$ and $\gamma > 0$ capture the adverse effects of harmful content and the opportunity costs of spending time on a platform.

Rational users maximize their utility, while naive users neglect the utility costs $-\delta h_p$ of consuming harmful content. We assume that all users choose the utility-maximizing engagement level $e_p^*(h_p) = \frac{(\eta_p - \theta_p)h_p + \theta_p}{2\gamma}$, which is independent of the utility costs $-\delta h_p$. Thus, the utility of a user who joins platform $p \in \{E, I\}$ is given by

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

The perceived utility of naive users is given by

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

For analytical simplicity, we set $\pi_p^t(x) = x$ for both $p \in \{E, I\}$ and $t \in \{n, r\}$. This means that the monetary value of one unit of engagement to a platform is identical for any user and both platforms.

We assume that the model's parameters are such that Assumption 1 is satisfied. For example, we assume that $\theta_p < \eta_p$ holds for either $p \in \{E, I\}$ to ensure that harmful content is more engaging. Further, we specify that $\theta_E < \theta_I$ and $\eta_E < \eta_I$ to model the incumbent's competitive advantage. Throughout the following analysis, we also focus on parameter combinations for which $V_E^n(1) < V_I^n(\check{h}_I)$, i.e., for which the incumbent enjoys a relatively large competitive advantage. Proposition 4 establishes that the naivety-focused equilibrium emerges for low values of ρ, the mixed-strategy equilibrium emerges for intermediate values of ρ, and the market dominance equilibrium emerges for high values of ρ.

The following figure visualizes which equilibria emerge in different parameter regions. We set $\theta_I = 3, \eta_I = 4$, $\gamma = 0.25$, $\delta = 20$, and $\eta_E = 3$. We consider different $\rho \in [0, 1]$ on the x-axis, and different $\theta_E \in [1, 2.75]$ on the y-axis.[16] Black points indicate that the naivety-focused equilibrium is the unique equilibrium (for the given parameter combination). Light grey points indicate that the mixed-strategy equilibrium characterized in Proposition 4 is the unique equilibrium. Dark grey points indicate that the market dominance equilibrium is the unique equilibrium.

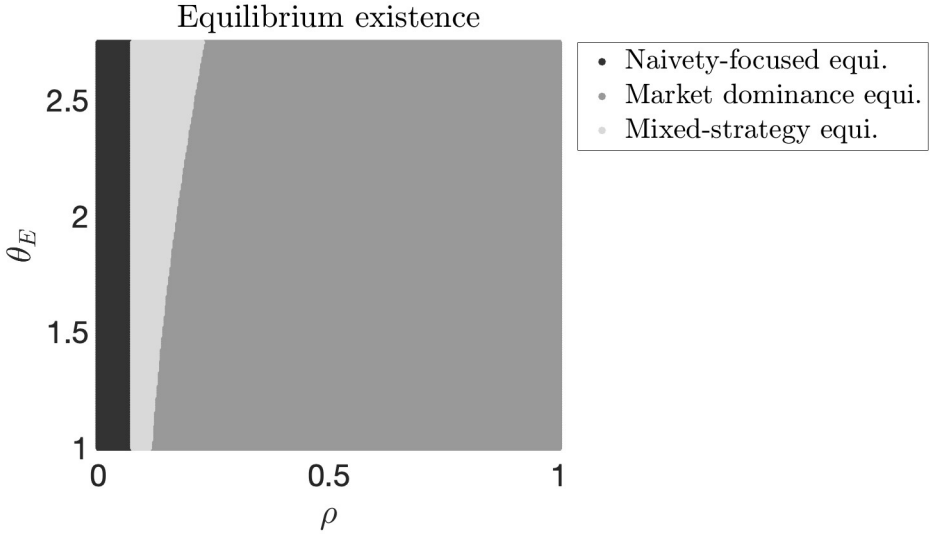

Figure 2: Equilibrium existence regions

[16] All parameter combinations we consider satisfy Assumption 1.

This graph visualizes the following facts: Firstly, the size of the parameter region for which the naivety-focused equilibrium exists is independent of the entrant's ability to generate utility for its users (as governed by the parameters θ_E and η_E). This is because this equilibrium exists if and only if $(1-\rho)e_I^*(1) \geq e_I^*(\tilde{h}_I)$, i.e., if the incumbent prefers to exclusively cater to naive users. Secondly, the parameter region for which the market dominance equilibrium exists shrinks when θ_E increases. Intuitively, this is because increases of θ_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.[17]

In the following, we consider the effects of two types of policy interventions. Firstly, we consider the effects of regulation that reduces the incumbent's competitive advantage by boosting the entrant's ability to generate utility for its users through the provision of non-harmful content. Secondly, we study initiatives that make the adverse effects of harmful content more salient. Within our model, these initiatives can be viewed as increases of ρ.

We begin by analyzing the effects of the first type of policy intervention. To do so, we set $\theta_I = 3$, $\eta_I = 4$, $\eta_E = 2.5$, $\gamma = 0.25$, $\delta = 20$, $\rho = 0.1$, and plot the expected harmful content shares either platform chooses in equilibrium, the incumbent's market share, and the expected (true) utilities which rational and naive users attain in equilibrium for different levels of $\theta_E \in [0.5, 1]$. Note that this expectation is formed over the harmful content shares chosen by platforms in a mixed-strategy equilibrium, and that the expected utilities of rational and naive users are the same in the market dominance equilibrium, so the corresponding points in the graph overlap for low values of θ_E.

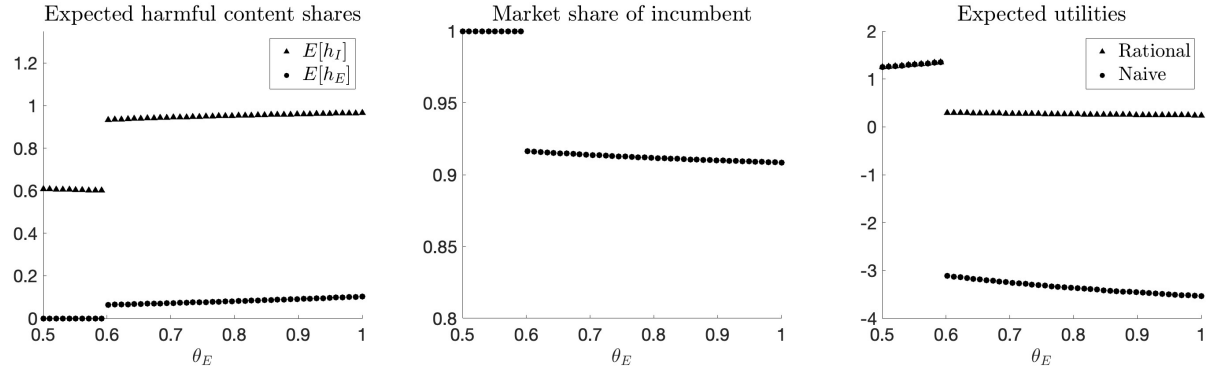

Figure 3: Comparative statics: Reductions of competitive advantage

This figure shows that increases of θ_E have non-monotonic effects on the expected amount of

[17]Formally, increases of θ_E lead to increases of $V_E(0)$, which means that $\check{h}_I$ decreases. This reduces the profits which the incumbent obtains in the market dominance equilibrium, and the deviation to $h_I = 1$ becomes more profitable.

harmful content that is displayed. To understand the visualized results, note that the market dominance equilibrium is played if $\theta_E < 0.6$ and that mixed-strategy equilibrium emerges if $\theta_E > 0.6$. In the market dominance equilibrium, increases of θ_E strengthen the competitive threat of the entrant, which induces the incumbent to lower the harmful content share it displays to retain rational users. Formally, $\check{h}_I$ falls when θ_E increases because $V_I(\check{h}_I) = V_E(0)$ holds by definition and $V_E(0)$ increases when θ_E increases. This benefits all users equally, since all users visit the incumbent.

When the market transitions from the market dominance equilibrium to the mixed-strategy equilibrium, there is an upward jump in the expected harmful content share chosen by both platforms. The underlying intuition is as follows: As θ_E crosses the threshold at $\theta_E \approx 0.6$, it becomes strictly more profitable for the incumbent to set $h_I = 1$ rather than $\check{h}_I$. Thus, the market dominance equilibrium ceases to exist, and the incumbent now displays the harmful content share $h_I = 1$ with strictly positive probability. Because the harmful content shares chosen by the platforms are strategic complements, this also induces the entrant to display more harmful content. At this point of discontinuity, the market share of the incumbent jumps down because a majority of rational users now visit the entrant. Moreover, the discontinuous upward jump in the amount of harmful content that is displayed goes along with a downward jump in users' expected utility.

In the mixed-strategy equilibrium, increases of θ_E raise the expected harmful content share chosen by both platforms. Intuitively, this is because an increase of θ_E allows the entrant to display more harmful content while holding the utility level it offers to any user fixed. Thus, the entrant will display more harmful content, which induces the incumbent to follow suit because the harmful content shares chosen by platforms are strategic complements. The increase in the expected harmful content shares chosen by both platforms leads to a reduction of users' expected utility, and particularly so for naive users who visit the incumbent and thus do not benefit directly from increases in θ_E.

In sum, these results suggest that regulation which improves the entrant's ability to generate utility for its users has non-monotonic effects: If the incumbent has a substantial competitive advantage, such regulation benefits users by strengthening the competitive threat the entrant poses. If the incumbent has a relatively weak competitive advantage, such pieces of legislation harm users by incentivizing both platforms to display more harmful content.

Having established this, we now consider the effects of the second type of policy intervention. To do so, we fix $\theta_I = 3$, $\theta_E = 1.5$, $\eta_I = 4$, $\eta_E = 2.5$, $\gamma = 0.25$, $\delta = 20$, and plot the expected harmful content shares either platform chooses in equilibrium, the incumbent's market share, and the expected true utilities of rational and naive users for $\rho \in [0, 0.3]$.

The expected utilities of rational and naive users are the same in the market dominance equilibrium, so the corresponding points in the graph overlap for high ρ.

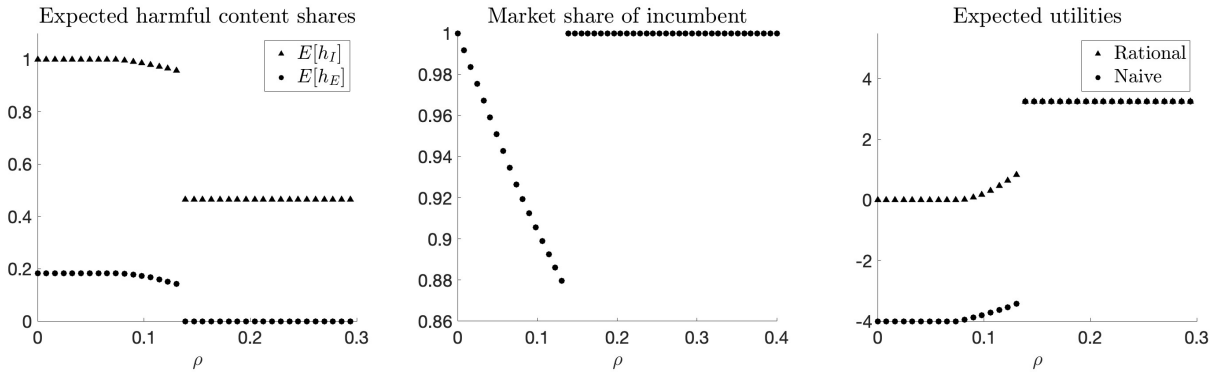

Figure 4: Comparative statics: ρ

This figure shows that increases in the share of rational users (ρ) have non-monotonic effects on the expected amount of harmful content that platforms display and the incumbent's market share. To understand the visualized results, note that the naivety-focused equilibrium is played if $\rho < 0.08$, the mixed-strategy equilibrium is played if $\rho \in (0.08, 0.14)$, and that the market dominance equilibrium is played if $\rho > 0.14$. In the naivety-focused equilibrium, the incumbent chooses $h_I^* = 1$ and the entrant chooses $h_E^* = \tilde{h}_E$. Increases of ρ thus only reduce the market share of the incumbent (since this platform is only visited by naive users).

In the mixed-strategy equilibrium, increases of ρ reduce the expected harmful content share both platforms display, and reduce the incumbent's market share. The expected harmful content share platforms display falls because it becomes relatively more profitable to choose low shares of harmful content when the share of rational users increases—this benefits users. The market share of the incumbent decreases because rational users predominantly visit the entrant (given that it displays much less harmful content). As the market transitions from the mixed-strategy equilibrium into the market dominance equilibrium (when $\rho \approx 0.14$), there is a downward jump in the expected harmful content share both platforms choose (which induces an upward jump in the expected utility of users), but the market share of the incumbent jumps up to one.

In sum, these results imply that initiatives which promote awareness regarding harmful content face a non-trivial trade-off: While they reduce the amount of harmful content that platforms display under competition, they may foster the monopolization of the market (if the entrant faces non-negligible fixed costs of being active), which harms users.

4 Leveling the Technological Playing Field

We now consider a benchmark in which the incumbent has no competitive advantage. The key take-away from this subsection is that, when platforms are symmetric, the user-optimal outcome—no exposure to harmful content—can emerge in equilibrium if the share of rational users is large enough. Formally, we consider a model that is analogous to the one described in Section 2, with a single exception: If both platforms set the same harmful content share, all users are indifferent between visiting either platform, i.e., $V_I(h) = V_E(h)$ and $V_I^n(h) = V_E^n(h)$ holds for all $h \in [0, 1]$. We refrain from providing a full equilibrium analysis of this setting, given that our focus is on settings with platform asymmetries, and only focus on establishing the aforementioned benchmark result:

Proposition 6 (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 an equilibrium in which* $h_I^* = h_E^* = 0$*.*

If the share of rational users is large enough, there exists an equilibrium in which all platforms display zero harmful content and rational users join either platform with equal probability. If any platform deviates from this equilibrium (by increasing its harmful content share), it will not be joined by rational users anymore. Thus, the most profitable deviation for any platform is to display maximal harmful content. Such a deviation is not profitable if the share of rational users is large enough, because its costs (in the form of significantly reduced demand) outweigh its benefits (higher engagement by naive users).

5 Policy Implications

Our analysis implies that the welfare effects of regulation in social media markets depend jointly on market structure and on platforms' incentives to choose harmful content shares. The key primitives are the incumbent's competitive advantage and the share of users who internalize the harms associated with harmful content. Because different instruments act on different margins, their effects are not interchangeable, and the appropriate policy mix depends on the market regime.

We distinguish three broad classes of regulatory instruments. A first reduces the incumbent's competitive advantage. A second reduces effective naivety by increasing the extent to which users internalize the harms associated with harmful content. A third directly constrains platforms' harmful content shares. These classes correspond closely to existing regulation: interoperability, data portability, and restrictions on self-preferencing; transparency,

warning-label, and choice-architecture interventions; and moderation or systemic-risk obligations directed at recommender systems.

Measures that reduce the incumbent's competitive advantage. A first class of interventions seeks to reduce the incumbent's advantage over rivals. In practice, this includes horizontal interoperability mandates, data portability requirements, and restrictions on conduct that entrenches the incumbent's installed base or data advantage. In the European context, natural examples are the GDPR portability right and the contestability provisions of the DMA, including rules aimed at limiting self-preferencing. In the model, these interventions reduce the utility gap the incumbent can sustain relative to the entrant for a given harmful content share. Technological change, including advances in AI, may also shift the same margin, but the direction of this effect is in general ambiguous.

Their effect depends on the share of users who internalize harms. In the market-dominance equilibrium, reducing the incumbent's advantage strengthens the entrant's threat and induces the incumbent to lower its harmful content share in order to retain users. By contrast, when the share of such users is too low, these measures need not improve welfare: the incumbent then caters only to naive users and chooses the maximal harmful content share, while a stronger entrant merely extracts more surplus from rational users. Even when locally beneficial, reducing the incumbent's advantage too aggressively may shift the market into the mixed-strategy region, where both platforms choose higher harmful content shares in expectation. Reductions in the incumbent's competitive advantage are therefore valuable only to the extent that they discipline harmful content choices.

Measures that reduce effective naivety. A second class of interventions aims to reduce effective naivety by increasing the extent to which users internalize the harms associated with harmful content. This includes warning labels, transparency requirements regarding recommender systems, digital-literacy initiatives, default time limits, break reminders, and related tools that help users regulate their own exposure. Examples include warning-label interventions such as California's Social Media Warning Law and, in the European context, the DSA's transparency requirements for recommender systems. In the model, such measures raise ρ, the share of rational users.

Their effect is to strengthen the responsiveness of demand to harmful content shares. As more users internalize the harms associated with harmful content, platforms face stronger pressure to moderate their feeds. This is why measures that reduce the incumbent's competitive advantage and measures that reduce effective naivety can work as complements: the former become effective only when enough users discipline platforms by responding to harmful-content choices. The appropriate form of these interventions depends on the source

of naivety. If naivety reflects imperfect awareness, transparency and information provision may suffice; if it instead reflects present bias, habit formation, or self-control problems, commitment devices or default protections become more relevant. The same logic extends to the alternative specification studied in Section C.4.

Content moderation measures. A third class of interventions directly constrains the harmful content shares that platforms may choose. In practice, this may take the form of duties to mitigate systemic risks from recommender systems, algorithmic auditing requirements, restrictions on specific categories of harmful content, or design obligations limiting the amplification of harmful material. The DSA provides a natural example through its systemic-risk assessment and mitigation framework for very large platforms. In the model, these interventions can be represented as a cap $\bar{h} \in (0, 1]$ such that each platform must choose $h_p \in [0, \bar{h}]$.

Proposition 7 (Content moderation and market dominance)**.**
Fix a moderation cap $\bar{h} \in (0, 1]$.

(i) If $\bar{h} \leq \check{h}_I$, then in every pure-strategy equilibrium, all users join the incumbent, and the incumbent chooses $h_I^* = \bar{h}$.

(ii) If $\bar{h} > \check{h}_I$, there exists a pure-strategy equilibrium in which $(h_I^*, h_E^*) = (\check{h}_I, 0)$ if and only if

$$V_E^n(\bar{h}) \leq V_I^n(\check{h}_I) \quad \textit{and} \quad \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I)) \geq (1-\rho)\,\pi_I^n(e_I^*(\bar{h})).$$

Content moderation does more than mechanically reduce harmful content shares. By truncating the race toward higher harmful content shares, it expands the set of parameters for which market dominance by the incumbent can be sustained. This matters because, in our framework, the market-dominance equilibrium yields higher user welfare than equilibria in which platforms segment the market and compete more aggressively for naive users. Moderation is also complementary to contestability measures: by ruling out very high harmful-content shares, it makes it less likely that a reduction in the incumbent's advantage shifts the market into the mixed-strategy region.

The role of AI. Generative AI may intensify the harm side of the model by making harmful, misleading, or emotionally charged content cheaper to produce, easier to personalize, and potentially harder for users to identify (Menczer et al., 2023). The effect of AI on the incumbent's competitive advantage is more nuanced. Following Gans (2024), the effect of AI in our setting depends on whether the data required to train and deploy AI systems can

move across firm boundaries. If data are portable, shared, or tradable, AI lowers the cost of producing high-quality recommendation, curation, and targeting systems, allowing entrants to narrow the incumbent's quality advantage. If data remain locked in, AI instead raises the return to the incumbent's proprietary user data and larger installed base, thereby reinforcing its advantage (Azoulay et al., 2024). The evidence in Brynjolfsson et al. (2025) is consistent with this interpretation, as it suggests that AI can codify and diffuse best practices, making some capabilities easier to replicate. From a policy perspective, this implies that regulation should address both the amplification of harmful AI-generated content and the competitive conditions under which AI capabilities are developed and deployed.[18]

The model also allows for targeted interventions when user groups differ systematically in the extent to which they internalize harms. Age is the most natural example. If younger users are less likely to internalize the harms associated with harmful content, stricter moderation rules, stronger default protections, or more restrictive recommender-system design for minors may be justified even when the corresponding interventions are less important for adults. Our framework also speaks to legislation restricting minors' access to social media, such as Australia's under-16 ban. If platforms can identify minors, such measures shut down the segment most likely to be characterized by a low share of rational users, improving welfare for that group without altering outcomes for adults. If age verification is imperfect, the effects are less clean, but removing some younger users may still raise the aggregate share of rational users and thereby shift the market toward the market-dominance equilibrium.[19]

The model also provides a useful lens through which to interpret cases of platform competition. A case such as Instagram versus TikTok, characterized by intense competition for engagement and a relatively young user base, may be closer to the mixed-strategy or naivety-focused region, in which reducing the incumbent's competitive advantage need not improve outcomes. By contrast, competition between $\mathbb{X}$ and entrants such as Meta Threads or BlueSky can be interpreted as a case in which some users with higher sensitivity to harmful content migrate to rival platforms, while the incumbent caters more strongly to users who remain relatively insensitive to such harms. These cases are only illustrative, but they clarify the broader lesson of the model: when the share of rational users is low, stronger competition alone need not discipline harmful content choices.

[18] In Europe, this concern is increasingly visible in current regulation. The AI Act's transparency rules for AI-generated or manipulated content become applicable on 2 August 2026, and the European Commission is currently facilitating the development of a Code of Practice on marking and labelling AI-generated content under Article 50.

[19] Implementation is challenging in practice because age-verification systems are imperfect and may create privacy and enforcement concerns. These frictions matter for the practical design of age-based regulation, but they do not affect the basic mechanism highlighted by the model.

Taken together, the analysis suggests a portfolio approach to regulation. Measures that reduce the incumbent's competitive advantage can improve outcomes, but only when enough users respond to harmful content shares for competition to discipline platforms. Measures that reduce effective naivety relax this constraint, while content moderation directly limits harmful content shares and makes reductions in the incumbent's competitive advantage safer to implement. In social media markets, competition policy and consumer-protection policy should therefore be designed jointly.

6 Robustness and Extensions

6.1 Multi-homing, network effects, and engagement differences

In Section C.1, we show that our results extend if users can multi-home. Specifically, we establish that all users must obtain weakly negative utility in any equilibrium with multi-homing, which implies that users are worse off in any such equilibrium than in the market dominance equilibrium. This is because the structure of any equilibrium with multi-homing (if it exists) is similar to the structure of the segmented equilibria from the baseline analysis: In any equilibrium with multi-homing, only naive users can multi-home, while rational users will all visit the same platform and do not multi-home.[20] One platform is thus only visited by naive users, which implies that it will choose the maximal harmful content share, thereby inducing its rival to set the harmful content share at which rational users obtain zero utility. In addition, the key properties of the equilibria from the baseline analysis are unaffected by the possibility of multi-homing, given that these properties follow from necessary conditions for the existence of these equilibria.

In Section C.2, we establish that our key predictions also emerge when there are network effects. We model the presence of network effects by allowing both the true and perceived utility of joining a given platform to be increasing in the measure of users who join this platform. The presence of network effects changes under what conditions equilibria exist—for example, the market dominance equilibrium becomes easier to sustain. However, it does not qualitatively change the predictions from the main analysis: The possible equilibrium candidates remain the same. Moreover, the welfare properties of these equilibria also carry over, given that these properties were based on necessary conditions for the existence of these equilibria, which also apply when there are network effects. In an equilibrium in which all users join the incumbent, for example, all users must obtain positive utility—otherwise, the

[20] There exists no equilibrium in which rational users multi-home, because one platform would prefer to deviate by slightly reducing its harmful content share to be visited by all rational users.

entrant could poach users even if users who join the incumbent benefit from network effects. In an equilibrium where all rational users visit a given platform and naive users visit its rival, the platform which naive users visit must choose the maximal harmful content share, and its rival will extract all surplus from the rational users who visit it.

In Section C.3, we consider a model in which rational and naive users on a given platform may have different engagement levels, which implies that rational and naive users may obtain different utilities on the same platform. This does not affect the welfare ordering of pure-strategy equilibria—in particular, we prove that the utility of all users must still be strictly larger in the market dominance equilibrium than in any other pure-strategy equilibrium. However, it may affect the welfare ordering of the market dominance equilibrium and mixed-strategy equilibria: Whereas rational users must obtain strictly larger utility in the market dominance equilibrium (holding their preferences fixed), this might not be true for naive users if the utility of rational and naive users on a given platform differs. Nevertheless, one can establish that user welfare jumps down as the market transitions from the market dominance equilibrium into the mixed-strategy equilibrium that emerges if the incumbent's competitive advantage is sufficiently strong. This is because naive users always visit the incumbent, so they are merely exposed to more harmful content and cannot be better off.

6.2 Different types of user heterogeneity

An alternative explanation of the fact that many users of social media do not seem to internalize the adverse effects thereof is that they form incorrect expectations about the share of harmful content platforms display. In Section C.4, we consider a model with such users. Specifically, users in this model variant are either rational or underestimate the share of harmful content any platform shows by a fixed factor. The key insights from our main analysis extend: In an equilibrium in which all users visit the incumbent, user welfare is strictly larger than in any other equilibrium. An analogue of the market dominance equilibrium emerges if the share of rational users is large.

In Section C.5, we consider a further dimension of consumer heterogeneity. Specifically, we consider a model in which all users are rational, but some are captive to the incumbent. This specification can capture other factors which keep users on dominant platforms such as the fear-of-missing-out, addiction, or switching costs. The predictions from the main analysis extend. Interestingly, an equilibrium in which all users visit the incumbent emerges if the share of captive users is small. This suggests that reductions of switching costs can increase the market share of the incumbent platform. The mechanics which underlie this result are as in the baseline analysis: If the share of captive users is large, the incumbent finds it

profitable to forgo non-captive users, who thus visit the entrant. If the share of captive users is small, the incumbent finds it profitable to ensure that it is visited by all users.

6.3 Continuous types

In Section C.6, we consider a variant of our model in which there is a continuum of user types who vary in the extent to which they internalize the adverse effects of harmful content. To begin, we consider a general version of this model and show that all users obtain strictly higher utility in any equilibrium in which all users visit the incumbent than in any other equilibrium. Thereafter, we analyze a parametric example using numerical analysis, and show that the relationship between user sophistication and the equilibria that emerge is as in our baseline model: For low (respectively, high) levels of sophistication, an analogue of the naivety-focused equilibrium (respectively, the market dominance equilibrium) emerges. Increases of user sophistication thus benefit users by reducing the share of harmful content platforms display, but may foster the monopolization of social media markets.

6.4 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: Firstly, all users must be strictly better off in any equilibrium in which all users visit the incumbent than in any other 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.

7 Conclusion

Social media platforms create substantial value, but they also generate significant harms for users' well-being. A central difficulty for policy is that these harms are not simply a consequence of market power in the usual sense. They arise in a business model in which platform profitability depends on capturing attention, and in which harmful content may be privately profitable even when it is socially costly.

Our analysis shows that the welfare effects of competition depend on the interaction between market structure and user sophistication and are non-monotonic. When the share of naive users is non-negligible, greater competition need not improve welfare and can even worsen outcomes, because it may intensify platforms' incentives to amplify harmful content. By contrast, policies that reduce behavioral biases or directly constrain harmful content amplification operate on a different margin and may improve welfare even when pro-competitive interventions alone do not.

Our analysis also helps clarify the logic of policy proposals aimed at changing the business model of digital platforms. Under the targeted-advertising business model, reaching the user-optimal outcome requires restrictive conditions: platforms must be nearly symmetric and the share of naive users must be negligible. This suggests that the distortion is not exhausted by market structure alone. In this environment, measures such as the digital advertising taxes proposed by Romer (2021) and Acemoglu and Johnson (2024) can be understood as operating on a deeper margin: by reducing the profitability of targeted advertising, they strengthen incentives to shift toward alternative revenue models, including subscription-based ones, in which profitability depends less on short-run attention and more on sustained user value and content quality.

A natural next step is to bring these mechanisms to the data. In companion work, we develop an empirical and experimental design aimed at measuring harmful-content exposure, user awareness, and contestability in real platform environments. More broadly, an important direction for future research is to study how platform incentives to amplify harmful content interact with network structure and with behavioral biases that are themselves shaped by platform design.

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 $h_I = 0$.

There exists no equilibrium in which the incumbent sets $h_I = 0$ with probability 1, which is a necessary condition for the implementation of the user-optimal outcome. To see this, note that rational users strictly prefer to visit the incumbent if it sets $h_I = 0$, no matter the entrant's strategy—this is because $V_I(0) > V_E(0)$ and $V_E(h_E)$ is decreasing in h_E. Thus, the incumbent can marginally increase its harmful content share without reducing its demand because the utility functions and the perceived utility functions are continuous (after the deviation, rational users still strictly prefer to visit the incumbent, while demand from naive users can only weakly increase). Thus, the deviation grants the incumbent higher engagement and is profitable. By implication, the equilibrium cannot exist. ■

Proof of Lemma 2:

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 play a pure strategy and all users of a given type (rational or naive) play the same 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. If platform l deviates by setting $h_l = 1$, this will raise the perceived 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. This means that they

would strictly prefer to join platform p, given that they would obtain strictly negative utility by joining platform l. Then, platform p would find it optimal to marginally increase the share of harmful content it displays—this is because the utility functions are continuous. 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 becomes even more profitable. Hence, the deviation is profitable, which is a contradiction.

Secondly, consider equilibria in which some users of a given type do not play the same strategy. This means that some user type must be indifferent in equilibrium. There exists no equilibrium in which all users are indifferent—if rational users are indifferent, this implies that $h_E^* < h_I^*$. But then, naive users cannot be indifferent because $V_E^n(h_E^*) < V_I^n(h_E^*) \leq V_I^n(h_I^*)$.

Suppose, for a contradiction, that there exists a pure-strategy equilibrium where some users visit the entrant and in which rational users are indifferent, while naive users are not. There exists no such equilibrium in which naive users visit the incumbent.[21] Suppose instead that naive users visit the entrant. If all rational users visit the incumbent, the entrant would set $h_E = 1$, so rational users cannot be indifferent. If some rational users visit the entrant, the incumbent would deviate from the equilibrium by slightly reducing its harmful content share—this is always possible, since rational users can only be indifferent if $h_I^* \geq \check{h}_I$. Thus, no such equilibrium exists.

Now consider a pure-strategy equilibrium in which naive users are indifferent, while rational users are not. Naive users visit the incumbent. If rational users also visit the incumbent, we are outside of the space of equilibria we consider in this part. Suppose instead that rational users visit the entrant. Then, $h_I^* = 1$ and $h_E^* = \tilde{h}_E$ must hold. In this equilibrium where some users visit the entrant, all users thus obtain weakly negative utility.

Part 2: In any 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 h_E in a small open interval above zero, given that $V_E(0) > 0$ and $V_E(h)$ is continuous. If the entrant sets such a harmful content share and is joined by rational users, it obtains strictly positive profits.

[21] If all rational users visit the entrant, the incumbent would set $h_I^* = 1$, so rational users cannot be indifferent. Suppose instead that some rational users join the incumbent. If $h_E^* > 0$, the entrant would prefer to deviate by slightly reducing h_E^*, which is profitable. If $h_E^* = 0$, then $h_I^* = \check{h}_I$ must hold. Then, all users would visit the incumbent, a contradiction to the premise.

Suppose, for a contradiction, that rational users join 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. ■

Proof of Proposition 1:

Part 1: In any equilibrium in which all users visit the incumbent with probability 1, the incumbent must choose $\check{h}_I$ with probability 1.

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 Γ_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 $V_I(h_I^*) > V_E(0)$. 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 (since $V_I(h_I)$ is continuous in h_I), 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 $V_I(h_I^*) < V_E(0)$. 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, $V_I(h_I^*) = V_E(0)$ must hold in equilibrium. This equation has a unique solution, namely $h_I^* = \check{h}_I$.

Part 2: There exists no equilibrium in which all users join the entrant.

Suppose, for a contradiction, that there exists an equilibrium in which all users join the entrant. Then, the incumbent would deviate by playing exactly the same strategy as the entrant. This deviation ensures that the incumbent obtains positive profits, and is thus profitable.[22]

Part 3: User welfare is 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.

Consider any MSE in which some users join the entrant with positive probability. Define $\bar{U}$ such that $V_I(\check{h}_I) = V_E(0) := \bar{U}$, where $\bar{U}$ is the utility which all users obtain in the pure-strategy equilibrium in which all users join the incumbent.

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 rather than the entrant or not joining any platform, no matter what h_E the entrant sets (since $V_I(h_I) > V_I(\check{h}_I) = V_E(0) \geq V_E(h_E)$ holds for any $h_I < \check{h}_I$ and any $h_E \in [0, 1]$). In a hypothetical equilibrium where the incumbent sets a $h_I' < \check{h}_I$, the incumbent would strictly prefer to deviate by setting a harmful content just above h_I'. This is because rational users still strictly prefer to visit the incumbent when it sets a harmful content share just above this h_I' by continuity of $V_I(h_I)$ (and the deviation cannot reduce demand from naive users but raises engagement). Thus, it can never be optimal for the incumbent to set a $h_I < \check{h}_I$.

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 $V_E(h_E) < \bar{U}$ must hold. For any $h_I > \check{h}_I$ such that $h_I \in \text{supp}\Gamma_I$, the inequality $V_I(h_I) < \bar{U}$ must hold. Both statements hold because $V_p(h_p)$ 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

[22] If Γ_E has an atom, this follows directly since the incumbent has a competitive advantage. Suppose alternatively that Γ_E has no atom. Then, the probability that the incumbent chooses a h_I below the h_E chosen by the entrant must be strictly positive, so the incumbent's demand is strictly positive.

$V_I(h_I) < \bar{U}$).[23] 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 entrant obtains must be strictly positive, since we consider a mixed strategy equilibrium where some users join the entrant with positive probability.[24] 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.

Part 4: User welfare is strictly larger in an equilibrium where all users visit the incumbent with probability 1 than in any other equilibrium.

By Lemma 2, user welfare is strictly larger in an equilibrium where all users visit the incumbent with probability 1 than in any other pure-strategy equilibrium. By parts 1-3, user welfare is strictly larger in an equilibrium where all users visit the incumbent with probability 1 than in any other mixed-strategy equilibrium. ■

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.

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 $V_E(0) = V_I(h_I^*)$.

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

[23] 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 $h_I < \check{h}_I$.

[24] If the entrant obtains zero demand for some $h_E \in \text{supp}\Gamma_E$, it must obtain zero demand for all $h_E \in \text{supp}\Gamma_E$ by the mixing indifference condition. But then, the entrant is not visited by any users in equilibrium.

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 $V_I(h_I^*) = V_E(0)$ must hold in equilibrium. To see this, note that $h_E = 0$ maximizes $V_E(h_E)$ by our assumptions. If $V_I(h_I^*) > V_E(0)$, the incentive constraint must be slack, a contradiction. If $V_I(h_I^*) < V_E(0)$, 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, given that the incentive constraint of rational users must bind and $V_E(h_E)$ is strictly decreasing.

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

Suppose, for a contradiction, that there exists an equilibrium in which all users join the entrant. 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 where some users of a given type join the incumbent and other users of the same type join the entrant, $h_I^* = \check{h}_I$ and $h_E^* = 0$ or $h_I^* = 1$ and $h_E^* = \tilde{h}_E$ must hold.

Consider an equilibrium in which some users of a given type join the entrant and others join the incumbent. By implication, users of this type must be indifferent between joining either platform. The arguments made in the proof of Lemma 2 imply that only naive users can be indifferent in equilibrium, while rational users must strictly prefer one platform over the other. If rational users visit the incumbent, $h_E^* = 0$ and $h_I^* = \check{h}_I$ must hold in equilibrium by the arguments in part 2. If rational users visit the entrant (and naive users visit the incumbent by our tie-breaking rule, given that they are indifferent), $h_E^* = \tilde{h}_E$ and $h_I^* = 1$ must hold by the arguments in part 1. ■

Proof of Proposition 3:

Part 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 negative utility by joining the incumbent, and zero utility by joining the entrant.

The entrant has no profitable deviations. By reducing h_E, it cannot attract more naive users (given that $V_E^n(1) < V_I^n(1)$ and $V_E^n(h_E)$ is weakly increasing), and will reduce engagement by rational users—this 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 $V_E^n(h_E)$ attains its maximum at $h_E = 1$ and $V_E^n(1) < V_I^n(h_I^*)$. 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. When deviating to any $h_I \in [0, \tilde{h}_I]$, the resulting profits are bounded from above by $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$. Under the assumption $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \leq (1-\rho)\pi_I^n(e_I^*(1))$, the equilibrium profits $(1-\rho)\pi_I^n(e_I^*(1))$ must thus be above the profits attainable through any such deviation as well.

Part 2: Definition of $\underline{\rho}$.

Define $\underline{\rho}$ such that $\underline{\rho}\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\underline{\rho})\pi_I^n(e_I^*(\tilde{h}_I)) = (1-\underline{\rho})\pi_I^n(e_I^*(1))$, which can be rewritten as:

$$\underline{\rho}\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\underline{\rho})\big(\underbrace{\pi_I^n(e_I^*(\tilde{h}_I)) - \pi_I^n(e_I^*(1))}_{<0}\big) = 0 \tag{11}$$

Note that the left-hand side of this equality is strictly increasing in ρ and equals 0 if $\rho = \underline{\rho}$. This implies that $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) < (1-\rho)\pi_I^n(e_I^*(1))$ holds for all $\rho < \underline{\rho}$. For all $\rho \in (0, \underline{\rho})$, the equilibrium thus exists.

Part 3: Equilibrium 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 (else, the incumbent would deviate by setting $h_I = 1$). 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 $h_I = 1$.[25] This is

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

a contradiction.

There exists no mixed-strategy equilibrium if $\rho < \underline{\rho}$. 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 not mix. ■

Proof of Proposition 4:

Part 1: Define $\bar{\rho}$ to satisfy $(1-\bar{\rho})\pi_I^n(e_I^*(1)) = \bar{\rho}\pi_I^r(e_I^*(\check{h}_I)) + (1-\bar{\rho})\pi_I^n(e_I^*(\check{h}_I))$. An equilibrium in which $h_I^* = \check{h}_I$ and $h_E^* = 0$ exists if and only if $\rho \geq \bar{\rho}$ and $V_E^n(1) \leq V_I^n(h_I^*)$ jointly hold.

Note that $(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$ holds if and only if $\rho \geq \bar{\rho}$.

An equilibrium with $h_I^* = \check{h}_I$ and $h_E^* = 0$ exists if $\rho \geq \bar{\rho}$ and $V_E^n(1) \leq V_I^n(h_I^*)$. In the proposed 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.

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 $V_E(h_E)$ 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 $V_E^n(1) > V_I^n(h_I^*)$. This holds because naive users strictly prefer to join the entrant after the deviation if $V_E^n(1) > V_I^n(h_I^*)$, 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

competitive advantage.

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

$$(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I)) \tag{12}$$

Also note that this equilibrium cannot exist if $\rho < \underline{\rho}$, because the incumbent would then deviate by setting $h_I^* = 1$. Similarly, the equilibrium cannot exist if $V_E^n(1) > V_I^n(\check{h}_I)$, because the entrant would deviate by setting $h_E^* = 1$.

Part 2: A unique mixed-strategy equilibrium exists if $V_E^n(1) < V_I^n(\check{h}_I)$ and $\rho \in (\underline{\rho}, \bar{\rho})$.

To begin, note that the following holds,

$$(1-\rho)\pi_I^n(e_I^*(1)) \in \left(\rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I)), \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))\right) \tag{13}$$

given that $\rho \in (\underline{\rho}, \bar{\rho})$

Assume further that $V_E^n(1) \leq V_I^n(\check{h}_I)$. Lemmas 3 - 6, which we present in the Online Appendix, establish that any mixed-strategy equilibrium must have the structure we described in the proposition.

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

$$\Pi_I(1) = \lim_{h_I \uparrow \tilde{h}_I} \Pi_I(h_I) \iff (1-\rho)\pi_I^n(e_I^*(1)) = \rho\lambda_E\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \tag{14}$$

$$\Pi_I(1) = \Pi_I(\underline{h}_I) \iff (1-\rho)\pi_I^n(e_I^*(1)) = \rho\pi_I^r(e_I^*(\underline{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\underline{h}_I)) \tag{15}$$

$$\Pi_E(\underline{h}_E) = \Pi_E(\tilde{h}_E) \iff \rho\pi_E^r(e_E^*(\underline{h}_E)) = \rho\lambda_I\pi_E^r(e_E^*(\tilde{h}_E)) \tag{16}$$

$$V_E(\underline{h}_E) = V_I(\underline{h}_I) \tag{17}$$

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

(i) A joint solution to equations (14) - (17) exists.

Firstly, note that there exists (under our assumptions) a unique $\lambda_E \in [0,1]$ that solves equation (14). To see this, note that left hand side of equation (14) is larger than the right hand side if $\lambda_E = 0$. In addition, the left hand side of equation (14) is smaller than the

right hand side if $\lambda_E = 1$ because $(1-\rho)\pi_I^n(e_I^*(1)) \leq \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$ holds by assumption. The fact that the right-hand side of equation (14) is continuous and strictly increasing in λ_E then implies the desired result.

Secondly, note that there always exists a unique λ_I such that equation (16) is satisfied.

Thirdly, there exists (under our assumptions) a unique $\underline{h}_I \in (\check{h}_I, \tilde{h}_I)$ that solves equation (15). 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 (15) is continuous and increasing in $\underline{h}_I$ then implies the desired result.

Fourthly, note that a unique solution $\underline{h}_E$ of equation (17) exists if $\underline{h}_I \geq \check{h}_I$ holds (we have verified the existence of an appropriate $\underline{h}_I$ in the last step). To see this, note that $V_I(\underline{h}_I) \in [0, V_I(\check{h}_I)]$ holds because $\underline{h}_I \geq \check{h}_I$. If $\underline{h}_E = 0$, the left-hand side of equation (17) is thus weakly larger than the right-hand side. If $\underline{h}_E = \tilde{h}_E$, the left-hand side of equation (17) is weakly smaller than the right-hand side. The fact that the left-hand side of (17) 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 (14) - (17) 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:

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

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:

$$\pi_E^r(e_E^*(r(h_I)))\rho\big[1 - F_I(h_I)\big] = \pi_E^r(e_E^*(\underline{h}_E))\rho, \tag{18}$$

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:

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

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:

$$(1-\rho)\pi_I^n(e_I^*(h_I)) + \rho\big(1 - F_E(r(h_I))\big)\pi_I^r(e_I^*(h_I))$$

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:

$$(1-\rho)\pi_I^n(e_I^*(r^{-1}(h_E))) + \rho\big(1 - F_E(h_E)\big)\pi_I^r(e_I^*(r^{-1}(h_E))) =$$

$$(1-\rho)\pi_I^n(e_I^*(\underline{h}_I)) + \rho\pi_I^r(e_I^*(\underline{h}_I)) \tag{19}$$

Solving for $F_E(h_E)$ yields:

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

$$\Longleftrightarrow$$

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

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. For the entrant, all $h_E < \underline{h}_E$ yield profits of $\rho\pi_E^r(e_E^*(h_E))$, which lie below $\rho\pi_E^r(e_E^*(\underline{h}_E))$. All $h_E > \tilde{h}_E$ yield zero profits because $V_E^n(1) \leq V_I^n(\check{h}_I)$.

For the incumbent, all $h_I < \underline{h}_I$ yield profits of $(1-\rho)\pi_I^n(e_I^*(h_I)) + \rho\pi_I^r(e_I^*(h_I))$, which lie below $(1-\rho)\pi_I^n(e_I^*(\underline{h}_I)) + \rho\pi_I^r(e_I^*(\underline{h}_I))$. All $h_I \in (\tilde{h}_I, 1)$ yield lower profits than $h_I = 1$. Thus, no platform has a profitable deviation.

Part 3: If $\rho \leq \underline{\rho}$ and $V_E^n(1) < V_I^n(\check{h}_I)$, there exists a unique equilibrium in which $h_I^* = 1$ and $h_E^* = \tilde{h}_E$.

Since $V_E^n(1) < V_I^n(\check{h}_I)$, there exists no pure-strategy equilibrium in which $h_E^* = 1$ and $h_I^* = \tilde{h}_I$, since the entrant would obtain zero profits in this equilibrium. Since $(1-\rho)\pi_I^n(e_I^*(1)) \geq \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$, given that $\rho \leq \underline{\rho}$, the market dominance equilibrium does not exist, since the incumbent would strictly prefer to deviate by setting $h_I = 1$, given that $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) > \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$.

If $\rho \leq \underline{\rho}$, there also exists no mixed-strategy equilibrium.

To see this, suppose firstly that $\rho < \underline{\rho}$, which implies that $(1-\rho)\pi_I^n(e_I^*(1)) > \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$. Then, the incumbent would strictly prefer to set $h_I = 1$ rather than any

other harmful content share in any equilibrium. But then, the entrant would strictly prefer to set $h_E = \tilde{h}_E$, so there cannot exist a mixed-strategy equilibrium.

If $\rho = \underline{\rho}$, $\underline{h}_I = \tilde{h}_I$ and $\lambda_E = 1$ must hold by equations (14) and (15). All naive users strictly prefer to visit the incumbent in equilibrium. Rational users must also visit the incumbent in equilibrium if it sets $\tilde{h}_I$, which must occur with positive probability (since we are in a mixed-strategy equilibrium). Otherwise, the incumbent would not be indifferent between setting $\tilde{h}_I$ and 1. But then, the entrant would prefer to deviate by setting a h_E just below $\tilde{h}_E$ that attracts all rational users. Thus, there exists no equilibrium in mixed strategies.

If $\rho \leq \underline{\rho}$, the naivety-focused equilibrium exists by Proposition 3. This establishes the result.

Part 4: If $\rho \in (\underline{\rho}, \bar{\rho})$ and $V_E^n(1) < V_I^n(\check{h}_I)$, there exists a unique equilibrium, namely the equilibrium in mixed strategies discussed above.

If $\rho \in (\underline{\rho}, \bar{\rho})$ and $V_E^n(1) < V_I^n(\check{h}_I)$, there exists no equilibrium in pure strategies. Since $V_E^n(1) < V_I^n(\check{h}_I)$, there exists no pure-strategy equilibrium in which $h_E^* = 1$ and $h_I^* = \tilde{h}_I$, since the entrant would obtain zero profits in this equilibrium. Since $(1-\rho)\pi_I^n(e_I^*(1)) > \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$, given that $\rho < \bar{\rho}$, the market dominance equilibrium does not exist, since the incumbent would prefer to deviate by setting $h_I = 1$. Since $(1-\rho)\pi_I^n(e_I^*(1)) < \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$, given that $\rho > \underline{\rho}$, the naivety focused equilibrium does not exist, since the incumbent would prefer to deviate by setting $h_I = \tilde{h}_I$.

The arguments in part 2 establish that there is a unique equilibrium in mixed strategies.

Part 5: If $\rho > \bar{\rho}$ and $V_E^n(1) < V_I^n(\check{h}_I)$, there exists a unique equilibrium in which $h_I^* = \check{h}_I$ and $h_E^* = 0$.

Since $V_E^n(1) < V_I^n(\check{h}_I)$, there exists no pure-strategy equilibrium in which $h_E^* = 1$ and $h_I^* = \tilde{h}_I$, since the entrant would obtain zero profits in this equilibrium. Since $(1-\rho)\pi_I^n(e_I^*(1)) < \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$, given that $\rho \geq \bar{\rho} > \underline{\rho}$, the naivety focused equilibrium does not exist, since the incumbent would prefer to deviate by setting $h_I = \tilde{h}_I$.

If $\rho > \bar{\rho}$, which implies that $(1-\rho)\pi_I^n(e_I^*(1)) < \rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I))$. In any equilibrium, the incumbent would thus never optimally set $h_I = 1$. But then, there exists no mixed-strategy equilibrium by Lemma 6. ■

Proof of Proposition 5:

Part 1: If $\rho < \underline{\rho}$, user welfare is identical under competition and in the monopoly benchmark.

Recall that $\underline{\rho}$ solves:

$$\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) = (1-\rho)\pi_I^n(e_I^*(1)) \tag{20}$$

For any $\rho < \underline{\rho}$, we thus have $\rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) < (1-\rho)\pi_I^n(e_I^*(1))$. Thus, the incumbent sets $h_I^* = 1$, both in the monopoly benchmark and in the competitive equilibrium. In the competitive equilibrium, the entrant sets $\tilde{h}_E$.

In the monopoly benchmark, rational users thus attain the utility zero, and naive users attain the utility $V_I(1)$. In the competitive equilibrium, rational users also attain zero utility, and naive users attain the utility $V_I(1)$. Thus, user welfare is identical.

Part 2: There exists a $\rho^m \in (\underline{\rho}, \bar{\rho})$ such that user welfare is strictly lower under competition if $\rho \in (\underline{\rho}, \rho^m)$.

Consider any $\rho \in (\underline{\rho}, \bar{\rho})$, and suppose that $V_E^n(1) < V_I^n(\check{h}_I)$. Then, the mixed-strategy equilibrium characterized earlier is the unique equilibrium. The equilibrium objects λ_I, λ_E, $\underline{h}_I$, and $\underline{h}_E$ must jointly solve the following set of equations:

$$(1-\rho)\pi_I^n(e_I^*(1)) = \rho\lambda_E\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) \tag{21}$$

$$(1-\rho)\pi_I^n(e_I^*(1)) = \rho\pi_I^r(e_I^*(\underline{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\underline{h}_I)) \tag{22}$$

$$\rho\pi_E^r(e_E^*(\underline{h}_E)) = \rho\lambda_I\pi_E^r(e_E^*(\tilde{h}_E)) \tag{23}$$

$$V_E(\underline{h}_E) = V_I(\underline{h}_I) \tag{24}$$

Under the assumptions that $\rho \in (\underline{\rho}, \bar{\rho})$ and $V_E^n(1) < V_I^n(\check{h}_I)$, there exists a unique joint solution $(\lambda_I, \lambda_E, \underline{h}_I, \underline{h}_E)$ to this set of equations—we have verified this in Proposition 4. The continuity assumptions in Assumption 1 imply that all these equations are continuous in the equilibrium objects and ρ. Thus, all equilibrium objects are continuous in ρ.

If $\rho \to \underline{\rho}$, this implies that $(1-\rho)\pi_I^n(e_I^*(1)) \to \rho\pi_I^r(e_I^*(\tilde{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I))$. By

equation (22), this implies that $\underline{h}_I \to \tilde{h}_I$, given that $e_I^*(h)$, $\pi_I^r(x)$, and $\pi_I^n(x)$ are all strictly increasing and continuous functions. By equation (24) and analogous arguments, this implies that $\underline{h}_E \to \tilde{h}_E$. By equation (23), this implies that $\lambda_I \to 1$. By equation (21), this implies that $\lambda_E \to 1$.

In the competitive equilibrium, the expected utility of naive users, who visit the incumbent, thus converges to $V_I(1) < 0$. The expected utility of rational users converges to zero. Thus, there is a ρ^m just above $\underline{\rho}$ such that user welfare is strictly negative for all $\rho \in (\underline{\rho}, \rho^m)$.

In the monopoly benchmark, all users attain the utility zero when $\rho > \underline{\rho}$. Thus, user welfare is strictly smaller under competition.

Part 3: If $\rho > \bar{\rho}$ and $V_E^n(1) < V_I^n(\check{h}_I)$, user welfare is larger in the competitive equilibrium than in the monopoly benchmark.

Suppose $\rho > \bar{\rho}$ and $V_E^n(1) < V_I^n(\check{h}_I)$. Given the definition of $\bar{\rho}$, this implies that:

$$(1-\rho)\pi_I^n(e_I^*(1)) \leq (1-\rho)\pi_I^n(e_I^*(\check{h}_I)) + \rho\pi_I^r(e_I^*(\check{h}_I)) \tag{25}$$

Under competition, there is thus a unique equilibrium in which $h_I^* = \check{h}_I$ and all users visit the incumbent. Under competition, all users thus attain utility $V_I(\check{h}_I)$.

In the monopoly benchmark, the incumbent sets $h_I = \tilde{h}_I$ if $\rho > \bar{\rho}$. To see this, note that this is optimal for the incumbent under monopoly because

$$(1-\rho)\pi_I^n(e_I^*(1)) \leq (1-\rho)\pi_I^n(e_I^*(\tilde{h}_I)) + \rho\pi_I^r(e_I^*(\tilde{h}_I)), \tag{26}$$

which holds true since $\check{h}_I < \tilde{h}_I$. In the monopoly benchmark, user welfare is thus zero. In the competitive benchmark, all users obtain strictly positive utility. Thus, user welfare is strictly higher under competition.

■

Proof of Proposition 6: Define $\bar{\rho}_I$ such that $(1-\bar{\rho}_I)\pi_I^n(e_I^*(1)) = 0.5\bar{\rho}_I\pi_I^r(e_I^*(0))$ and $\bar{\rho}_E$ such that $(1-\bar{\rho}_E)\pi_E^n(e_E^*(1)) = 0.5\bar{\rho}_E\pi_E^r(e_E^*(0))$.

Set $\rho' = \max\{\bar{\rho}_I, \bar{\rho}_E\}$. If $\rho > \rho'$, $(1-\rho)\pi_p^n(e_p^*(1)) < 0.5\rho\pi_p^r(e_p^*(0))$ holds for both $p \in \{E, I\}$. Then, there exists an equilibrium in which $h_E^* = 0$ and $h_I^* = 0$ and rational users visit either platform with probability 0.5.

In equilibrium, either platform's profits are weakly greater than $0.5\rho\pi_p^r(e_p^*(0))$. If it deviates (by increasing h_p), it will not be visited by rational users anymore. Thus, deviation

profits are weakly smaller than $(1-\rho)\pi_p^n(e_p^*(1))$. Since $\rho > \rho'$, there cannot be a profitable deviation, and the equilibrium exists. ■

Proof of Proposition 7:

Part 1: If $\bar{h} \leq \check{h}_I$, then in every pure-strategy equilibrium all users join the incumbent, and the incumbent chooses $h_I^* = \bar{h}$.

Suppose $\bar{h} \leq \check{h}_I$. By definition, $V_I(\check{h}_I) = V_E(0)$ and $V_E(\cdot)$ is strictly decreasing. Hence, for any $h_E \in [0, \bar{h}]$,

$$V_I(\bar{h}) \geq V_I(\check{h}_I) = V_E(0) \geq V_E(h_E),$$

with strict inequality if $\bar{h} < \check{h}_I$. Thus, if the incumbent sets $h_I = \bar{h}$, rational users weakly prefer the incumbent for any h_E. Naive users also strictly prefer the incumbent given that perceived utility is weakly increasing in h_I and the incumbent has a competitive advantage.

In any pure-strategy equilibrium, the incumbent must therefore choose $h_I^* = \bar{h}$: any lower h_I would leave demand unchanged while reducing engagement and profits. It remains to rule out equilibria in which some rational users join the entrant when $\bar{h} = \check{h}_I$. But in such a case, the incumbent can profitably deviate to some $h_I' < \check{h}_I$ arbitrarily close to $\check{h}_I$, thereby attracting all rational users while losing only an arbitrarily small amount of engagement (naive users still strictly prefer to visit the incumbent after the deviation, since they would strictly prefer to visit the incumbent if both platforms choose the harmful content share $\bar{h}$). Hence, in every pure-strategy equilibrium all users join the incumbent and $h_I^* = \bar{h}$.

Part 2: Existence of the market dominance equilibrium.

Consider the profile $(h_I^*, h_E^*) = (\check{h}_I, 0)$. By definition of $\check{h}_I$, we have

$$V_I(\check{h}_I) = V_E(0),$$

so rational users are indifferent between the two platforms and may all join the incumbent in equilibrium. Naive users join the incumbent if and only if the entrant cannot make itself more attractive to them by choosing some feasible harmful-content share. Since $V_E^n(h_E)$ is weakly increasing in h_E, the entrant's most attractive choice for naive users is $h_E = \bar{h}$. Hence, naive users join the incumbent if

$$V_E^n(\bar{h}) \leq V_I^n(\check{h}_I).$$

Given this condition, the entrant has no profitable deviation: any $h_E > 0$ makes it strictly less attractive to rational users than $h_E = 0$, and no feasible h_E attracts naive users. For the incumbent, any deviation to $h_I < \check{h}_I$ is unprofitable because it leaves demand unchanged while reducing engagement. Any deviation to $h_I > \check{h}_I$ loses rational users, so the most profitable such deviation is to $h_I = \bar{h}$, which leaves the incumbent with only naive users. This deviation is unprofitable if and only if

$$\rho\pi_I^r(e_I^*(\check{h}_I)) + (1-\rho)\pi_I^n(e_I^*(\check{h}_I)) \geq (1-\rho)\pi_I^n(e_I^*(\bar{h})).$$

Thus, $(h_I^*, h_E^*) = (\check{h}_I, 0)$ is a pure-strategy equilibrium if and only if the two stated inequalities jointly hold. ■

References

D. Acemoglu and S. Johnson. The urgent need to tax digital advertising. Network Law Review, Spring 2024, 2024. URL `https://www.networklawreview.org/acemoglu-johnson/`. Accessed 24 March 2026.

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):629–676, 2020.

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

A. Ambrus, E. Calvano, and M. Reisinger. Either or both competition: A "two-sided" theory of advertising with overlapping viewerships. American Economic Journal: Microeconomics, 8(3):189–222, 2016.

S. P. Anderson and S. Coate. Market provision of broadcasting: A welfare analysis. The review of Economic studies, 72(4):947–972, 2005.

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.

P. Azoulay, J. L. Krieger, and A. Nagaraj. Old moats for new models: Openness, control, and competition in generative ai. NBER Working Paper 32474, National Bureau of Economic Research, May 2024. URL https://www.nber.org/papers/w32474.

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.

E. Brynjolfsson, D. Li, and L. R. Raymond. Generative ai at work. The Quarterly Journal of Economics, 140(2):889–942, May 2025. doi: 10.1093/qje/qjae044. URL https://academic.oup.com/qje/article/140/2/889/7990658.

L. Bursztyn, B. Handel, R. Jiménez-Durán, and C. Roth. When product markets become collective traps: The case of social media. American Economic Review, 115(12):4105–4136, 2025.

B. Caillaud and B. Jullien. Chicken & egg: Competition among intermediation service providers. RAND journal of Economics, pages 309–328, 2003.

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

M. J. Dreier, S. I. Boyd, S. L. Jorgensen, R. Merai, J. Fedor, K. C. Durica, C. A. Low, and J. L. Hamilton. Adolescents' daily social media use and mood during the covid-19 lockdown period. Current Research in Ecological and Social Psychology, 7:100196, 2024.

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

European Commission. Commission preliminarily finds tiktok's addictive design in breach of the digital services act. 2026a.

European Commission. Commission launches investigation into shein under the digital services act. 2026b.

X. Gabaix and D. Laibson. Shrouded attributes, consumer myopia, and information suppression in competitive markets. The Quarterly Journal of Economics, 121(2):505–540, 2006.

J. S. Gans. Market power in artificial intelligence. NBER Working Paper 32270, National Bureau of Economic Research, Mar. 2024. URL `https://www.nber.org/papers/w32270`.

E. Giovannetti and P. Siciliani. Platform competition and incumbency advantage under heterogeneous lock-in effects. Information Economics and Policy, 63:101031, 2023. ISSN 0167-6245. URL `https://www.sciencedirect.com/science/article/pii/S0167624523000161`.

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.

A. Hagiu and B. Jullien. Search diversion and platform competition. International Journal of Industrial Organization, 33:48–60, 2014.

P. Heidhues and B. Kőszegi. Naivete-based discrimination. The Quarterly Journal of Economics, 132(2):1019–1054, 2017.

P. Heidhues, B. Kőszegi, and T. Murooka. Exploitative innovation. American Economic Journal: Microeconomics, 8(1):1–23, 2016.

R. Hoong. The behavioural economics of social media: A study of self commitment devices and the facebook privacy paradox, 2019.

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.

D.-S. Jeon and P. Rey. Platform competition and innovation. Technical Report 24-1566, Toulouse School of Economics, 2026. TSE Working Paper, September 2024; revised February 2026.

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.

L. Marciano, C. C. Driver, P. J. Schulz, and A.-L. Camerini. Dynamics of adolescents' smartphone use and well-being are positive but ephemeral. Scientific reports, 12(1):1316, 2022.

F. Menczer, D. Crandall, Y.-Y. Ahn, and A. Kapadia. Addressing the harms of ai-generated inauthentic content. Nature Machine Intelligence, 5:679–680, 2023. doi: 10.1038/s42256-023-00690-w.

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

G. G. Parker and M. W. Van Alstyne. Two-sided network effects: A theory of information product design. Management science, 51(10):1494–1504, 2005.

M. Peitz and T. M. Valletti. Content and advertising in the media: Pay-tv versus free-to-air. international Journal of industrial organization, 26(4):949–965, 2008.

PEW Research, 2025. URL https://www.pewresearch.org/internet/2025/04/22/teens-social-media-and-mental-health/#:~:text=Still%2C%20teens%20are%20growing%20more,they%20negatively%20affect%20them%20personally.

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

M. Risco and M. Lleonart-Anguix. Feed for good? on the effects of personalization algorithms in social platforms. Technical Report 580, CRC TR 224 Discussion Paper Series, 2024.

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

P. Romer. Digital ad-tax. https://adtax.paulromer.net/, 2021. Accessed 24 March 2026.

C. Sagioglou and T. Greitemeyer. Facebook's emotional consequences: Why facebook causes a decrease in mood and why people still use it. Computers in Human Behavior, 35:359–363, 2014.

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

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.

Addictive Platform Design: Competition, Awareness, and Regulation

Online Appendix

B Auxiliary lemmata characterizing mixed-strategy equilibria

Lemma 3. *Consider any MSE in which there exists a* $h_j' \in \text{supp } \Gamma_j \setminus 1$. *Then* Γ_{-j} *must have an atom at* $\bar{h}_{-j}$ *or an atom at* 1.

Proof of Lemma 3 : Consider any MSE in which there exists a $h_j' \in \text{supp}\Gamma_j \setminus 1$ and suppose, for a contradiction, that Γ_{-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 $V_j(\bar{h}_j) = V_{-j}(\bar{h}_{-j})$ must hold). 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 4. *Consider any MSE in which there exists a* $h_E \in supp\Gamma_E \backslash 1$ *and a* $h_I \in supp\Gamma_I \backslash 1$. *Then,* $\bar{h}_E = \tilde{h}_E$ *and* $\bar{h}_I = \tilde{h}_I$ *must hold. For both* $j \in \{E, I\}$, *the distribution* Γ_j *must be atomless and gapless on* $\left[\inf \text{supp}\Gamma_j, \tilde{h}_j\right)$.

Proof of Lemma 4 :

(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 λ_{-j} as the probability that platform $-j$ sets $h_{-j} = \bar{h}_{-j}$ or $h_{-j} = 1$.

Suppose Γ_{-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 Γ_{-j} has an atom at $\bar{h}_{-j}$. Then, Γ_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 3, it follows that $\bar{h}_j = \tilde{h}_j$ must hold, since $V_j(\bar{h}_j) = V_{-j}(\bar{h}_{-j})$ must hold in equilibrium. This is a contradiction.

(ii) The distributions must be atomless.

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

Lemma 5. *Consider any MSE in which there exists a $h_E \in supp\Gamma_E \setminus 1$ and a $h_I \in supp\Gamma_I \setminus 1$. In equilibrium, $V_E(\underline{h}_E) = V_I(\underline{h}_I)$ must hold.*

Proof of Lemma 5 : Suppose, for a contradiction, that $V_j(\underline{h}_j) < V_{-j}(\underline{h}_{-j})$ 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 6. *If $V_I^n(\tilde{h}_I) > V_E^n(1)$, 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 4 - 5. Moreover, $1 \in \text{supp}\Gamma_I$ and $\tilde{h}_E \in \text{supp}\Gamma_E$ must hold.

Proof of Lemma 6:

(i) If $V_I^n(\tilde{h}_I) > V_E^n(1)$, 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 incumbent. Thus, the entrant would obtain zero demand in equilibrium, a contradiction.

(ii) If $V_I^n(\tilde{h}_I) > V_E^n(1)$, 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 4 and previous arguments, $\tilde{h}_I \in \Gamma_I$ must hold. Since $1 \in \text{supp}\Gamma_E$, Γ_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 $V_I^n(h_I') = V_E^n(1)$. Suppose $\underline{h}_I < h_I'$. Then, the distribution Γ_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 $V_I^n(h_I) > V_I^n(h_I') = V_E^n(1)$ holds for almost all $h_I \in \text{supp}\Gamma_I$ and since the distribution Γ_I cannot have an atom at $\underline{h}_I$. This is a contradiction.

(iii) If $V_I^n(\tilde{h}_I) > V_E^n(1)$, any mixed-strategy equilibrium must satisfy the properties laid out in Lemmas 4 - 5. 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 3 establishes that Γ_E must have an atom at $\tilde{h}_E$, since $1 \notin \Gamma_E$. Thus means that Γ_I cannot have an atom at $\tilde{h}_I$, so it must have an atom at 1. ■

C Extensions

C.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:

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

where $g_I = 1 - h_I$ and $g_E = 1 - h_E$.

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:

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

The utility which users obtain when joining a platform p is as in the parametric example of Section 4.2. The perceived utility which naive users obtain when joining a platform p is as in the parametric example of Section 4.2. 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.

C.2 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

$$V_p(h_p, s_p) = \frac{(h_p\eta_p(s_p) + (1-h_p)\theta_p(s_p))^2}{4\gamma} + (1-h_p) - \delta h_p, \tag{C.3}$$

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

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

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 $s \in [0,1]$, $V_I(h,s) > V_E(h,s)$ and $V_I^n(h,s) > V_E^n(h,s)$ hold.*
2. *For both $p \in \{I, E\}$ and any $s_p \in [0,1]$, the function $V_p(h, s_p)$ strictly decreases in h, $V_p(1, s_p) < 0$ holds, and $V_p(0, s_p) > 0$ holds.*
3. *For both $p \in \{I, E\}$, $\frac{\partial V_p^n(h,s_p)}{\partial h} > 0$ 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 9 (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.

C.3 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 10 (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. Our first main result, 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.

C.4 An alternative definition of naivete

In this subsection, we consider an extension in which the behavior of naive users is modeled slightly differently. As before, we suppose that users' engagement on a platform p is represented by the function $e_p^*(h_p)$ and denote the true utility which users attain when they join platform p as $V_p(h_p)$. Rational users maximize $V_p(h_p)$ by choosing which platform to join, while naive users maximize the perceived utility $V_p^n(h_p)$, where

$$V_p^n(h_p) := V_p(\alpha h_p), \tag{C.5}$$

and $\alpha \in (0,1)$ is an exogenously given constant. Under this definition of naivete, naive users can be understood as users who underestimate the share of harmful content which a platform displays by the factor α. We impose the same assumptions on the functions $e_p^*(h_p)$ and $V_p(h_p)$ as in the main analysis, namely that:

Assumption 3. *The following assumptions hold:*

1. *$V_I(h) > V_E(h)$ holds for all $h \in [0,1]$.*
2. *For both $p \in \{I, E\}$, $V_p(h)$ is continuously differentiable and strictly decreasing in h. Further, $V_p(1) < 0 < V_p(0)$ and $V_p^n(1) < 0$ holds.*

3. For both $p \in \{I, E\}$, $e_p^*(h)$ is strictly increasing in h.

Everything else is as in the baseline model. We refer to this model as the *misperception model*.

To begin the formal analysis, we define two critical levels of harmful content shares, namely $\check{h}_I$ and h'_I. These objects satisfy:

$$V_E(0) = V_I(\check{h}_I) \quad ; \quad V_E^n(0) = V_I^n(h'_I) \tag{C.6}$$

Note that $V_E(0) = V_E^n(0)$ and that $\check{h}_I < h'_I$ hold by construction. Intuitively, $\check{h}_I$ is the harmful content share at which *rational users* are indifferent between visiting the entrant and the incumbent if the entrant displays no harmful content and the incumbent chooses the harmful content share $\check{h}_I$. Moreover, h'_I is the harmful content share at which *naive users* are indifferent between visiting the entrant and the incumbent if the entrant displays no harmful content and the incumbent chooses the harmful content share h'_I.

We begin by characterizing the possible candidates for a pure-strategy equilibrium.

Lemma 7. *If a pure-strategy equilibrium exists, then $h_I^* = \check{h}_I$ and $h_E^* = 0$ must hold and all users join the incumbent in equilibrium. This equilibrium exists if and only if $(1 - \rho)\pi_I^n(e_I^*(h'_I)) \leq \rho\pi_I^r(e_I^*(\check{h}_I)) + (1 - \rho)\pi_I^n(e_I^*(\check{h}_I))$.*

The only pure-strategy equilibrium that can exist within the misperception model is thus analogous to the market dominance equilibrium from the baseline analysis: All users visit the incumbent in equilibrium, the entrant displays no harmful content, and the incumbent displays a relatively low share of harmful content that barely retains rational users.

Further, our prediction regarding the ordering of user welfare across equilibria extends:

Proposition 11. *Consider the misperception model. In any equilibrium in which all users visit the incumbent, all users obtain a weakly higher utility than in any other equilibrium, and a strictly positive measure of users attain a strictly higher utility.*

The intuition underlying this result is analogous to the familiar logic from the baseline model: In any equilibrium in which all users visit the incumbent, the incumbent sets the relatively low harmful content share $\check{h}_I$ because competition is centered around rational users. In any other equilibrium, the incumbent prefers to forego some rational users in order to obtain higher engagement from naive users (by setting larger harmful content shares). Thus, users are worse off than in an equilibrium in which all users visit the incumbent.

Taken together, the preceding analysis implies that other key predictions from the baseline analysis also extend: Firstly, user migration from a dominant platform to a smaller

entrant platform reduces user welfare by incentivizing the incumbent platform to display more harmful content. Secondly, increases in the share of rational users can raise user welfare by reducing the share of harmful content platforms display. However, the incumbent will attain an even more dominant position in this process.

C.5 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 ρ 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. We directly present the main equilibrium characterization.

Proposition 12 (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 ρ^1 and ρ^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 $\rho > \rho^2$.*

C.6 Continuous types

Overview: In this section, we consider a variant of our model in which there is a continuum of user types who vary in the extent to which they internalize the adverse effects of harmful content. To begin, we consider a general version of this model and show that all users obtain strictly higher utility in any equilibrium in which all users visit the incumbent than in any other equilibrium. Our insight that user migration away from a dominant platform can harm users by inducing platforms to display more harmful content thus extends readily.

Thereafter, we study a parametric example and show (using numerical analysis) that the relationship between user sophistication and the equilibria that emerge is as in our baseline model: For low (respectively, high) levels of sophistication, an analogue of the naivety-focused equilibrium (respectively, the market dominance equilibrium) emerges. As in our baseline model, increases of user sophisticaton thus benefit users by reducing the share of harmful content platforms display, but may foster the monopolization of social media markets.

Model: There is a unit mass of users. The utility of any user who visits platform p is

$$U_p(h_p) = b_p(h_p) - c(h_p),$$

where $b_p(h_p)$ and $c(h_p)$ are increasing functions. We normalize $c(0) = 0$.

The degree to which consumers internalize the costs of harmful content is given by their type $t_i \in T$, where $t_i \sim H$. The perceived utility a user with type t_i attains by visiting platform p is

$$U_p^s(h_p, t_i) = b_p(h_p) - t_i c(h_p), \tag{C.7}$$

where h_p is the harmful content share chosen by the platform and the superscript "s" indicates that this is a *subjectively* evaluated utility. We assume that $1 \in \text{supp}T$. A user with type $t_i = 1$ is fully rational, i.e., $U_p^s(h_p, 1) = U_p(h_p)$

If the incumbent chooses the harmful content share h_I and the entrant chooses h_E, a user with type t visits the incumbent if and only if

$$b_I(h_I) - tc(h_I) \geq b_E(h_E) - tc(h_E).$$

Everything else is as in the baseline model. Furthermore, we impose the following assumption which ensures that the response of any given user to more harmful content is globally similar.

Assumption 4. *For any $t_i \in T$ and either $p \in \{E, I\}$, the function $U_p^s(h_p, t_i)$ is either*

globally strictly decreasing or weakly increasing in h_p. Furthermore, any user's engagement level is increasing in harmful content.

We integrate the incumbent's competitive advantage by imposing the following assumption.

Assumption 5. *For any $t_i \in T$ and $h \in [0,1]$, $U_I^s(h,t_i) > U_E^s(h,t_i)$ holds. Moreover, if $U_I^s(h_I,t_i)$ is strictly decreasing in h_I for some t_i, then $U_E^s(h_E,t_i)$ is also strictly decreasing.*

This assumption reflects that a given piece of harmful content either generates more utility for a user or generates less utility cost if it is displayed by the incumbent. This is equivalent to the incumbent enjoying a competitive advantage.

General results:

We begin by establishing that our first key result from the baseline analysis carries over to the settings we consider in this extension:

Proposition 13. *In any equilibrium in which all users visit the incumbent with probability 1, all users obtain strictly higher utility than in any other equilibrium.*

The intuition which underlies this result is familiar: In any equilibrium in which the incumbent is visited by all users, all users must obtain strictly positive utility. Otherwise, the entrant could poach a strictly positive measure of users—formally, this statement holds because $1 \in \text{supp}T$ and the subjective utility functions are continuous. Moreover, the incumbent will set the harmful content share $\check{h}_I$ with probability 1 in such an equilibrium, where $\check{h}_I$ is as defined in the main analysis and solves $U_I^s(\check{h}_I,1) = U_E^s(0,1)$.

In any equilibrium in which the entrant is visited by some users, the incumbent will never set a harmful content share below $\check{h}_I$ and will set a harmful content share above $\check{h}_I$ with strictly positive probability. To see why this holds true, note that it is never optimal for the incumbent to set a $h_I < \check{h}_I$. In an equilibrium in which some users visit the entrant, the incumbent must set a harmful content share strictly above $\check{h}_I$ with positive probability—if it sets $\check{h}_I$ with probability 1, all users would visit the incumbent. Because the true utility of any user is decreasing in harmful content and the incumbent has a competitive advantage, the true utility of all users is thus strictly lower than $U_I^s(\check{h}_I,1)$ in such an equilibrium.

Numerical analysis:

Setup: Now, we build further intuition by considering a particular parametric version of the

model we just laid out. Specifically, suppose that $b_p(h_p) = \alpha_p + \beta_p h_p$, and that $c(h_p) = h_p$. For simplicity, we also impose that $e_p^*(h_p) = h_p$ for both platforms p.

We further set $\alpha_I = 0.25$, $\beta_I = 0.5$, and $\beta_E = 0.4$, and consider two different $\alpha_E \in \{0.05, 0.2\}$ to model different levels of the competitive advantage. Lower levels of α_E signify that the incumbent has a stronger competitive advantage.

We assume that the users' types are uniformly distributed on the interval $[\kappa, 1]$, where $\kappa \in (0, 1)$ is a parameter and governs the overall level of user sophistication. Increases of κ signify that there are more users that internalize the adverse effects of harmful content to a significant extent—thus, increases of κ capture increases in the level of user sophistication (as increases in ρ do within our baseline model).

The expected revenue of either platform $p \in \{E, I\}$ is denoted by $\Pi_p(h_I, h_E)$, where:

$$\Pi_I(h_I, h_E) = \int_\kappa^1 \mathbb{1}[\alpha_I + (\beta_I - t_i)h_I > \alpha_E + (\beta_E - t_i)h_E](h_I)(1/(1-\kappa))\mathrm{dt_i}$$

$$\Pi_E(h_I, h_E) = \int_\kappa^1 \mathbb{1}[\alpha_I + (\beta_I - t_i)h_I < \alpha_E + (\beta_E - t_i)h_E](h_E)(1/(1-\kappa))\mathrm{dt_i}$$

We restrict attention to equilibria in pure strategies. A pair (h_I^*, h_E^*) is an equilibrium if and only if $h_E^* = \mathrm{arg}max_{h_E}\Pi_E(h_I^*, h_E)$ and $h_I^* = \mathrm{arg}max_{h_I}\Pi_I(h_I, h_E^*)$.

Equilibrium characterization: We now plot the equilibrium levels of h_I^* and h_E^* that emerge for different levels of $\kappa \in [0, 1]$ (which are plotted on the x-axis) and different levels of $\alpha_E \in \{0.05, 0.2\}$ (which are held fixed in each graph):

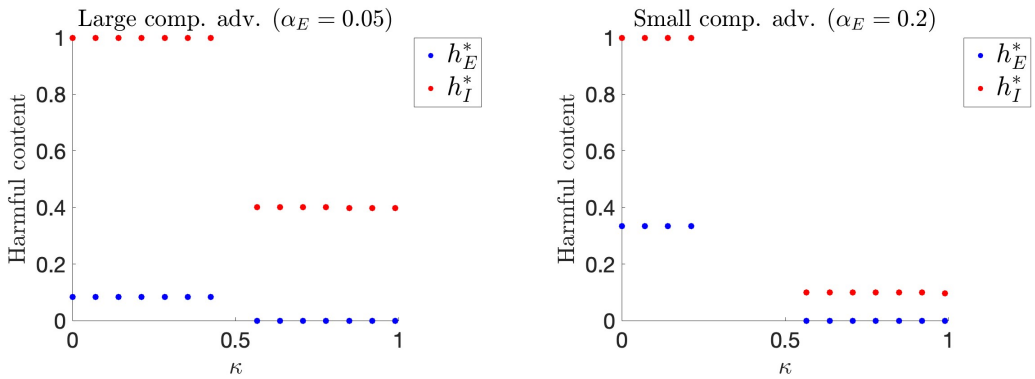

Figure C.1: Continuous users: Equilibrium outcomes

If the degree of user sophistication is small (i.e., if κ is small), an analogue of the naivety-focused equilibrium emerges. In equilibrium, the incumbent sets $h_I^* = 1$ and the entrant sets

the harmful content share at which a fully rational user is indifferent between joining the entrant and not joining any platform—this is just the harmful content share $\tilde{h}_E$ from the baseline analysis.[26]

If the degree of user sophistication is large (i.e., if κ is large), the incumbent sets a harmful content share that guarantees that it is visited by users that are relatively sophisticated. In particular, it sets the harmful content share $\check{h}_I$ which makes a fully rational user indifferent between visiting the entrant and the incumbent if the entrant displays no harmful content—this is just the harmful content share $\check{h}_I$ from the baseline analysis. If the incumbent sets this harmful content share and user sophistication is high, the entrant can never attract any users and optimally displays no harmful content.

One can verify that the market share of the incumbent is one if $\check{h}_I = 1$ and $h_E^* = 0$. We also note that no equilibrium in pure strategies exists if κ is at intermediate levels, i.e., when user sophistication is neither very weak nor very strong—this outcome mirrors the predictions we obtained in the baseline analysis. Finally, our numerical analysis suggests that the equilibria we have found are unique (for a given parameter combination).

As in the baseline analysis, increases of user awareness thus benefit users by reducing the share of harmful content platforms display, but may promote the monopolization of social media markets. The intuition is familiar: The incumbent prioritizes naive users because these are more profitable. Increases in user sophistication thus only affect whether users with relatively large levels of sophistication visit the incumbent. Because these users only visit the incumbent if it displays a relatively low share of harmful content, reductions of the harmful content platforms display (induced by increases in user sophistication) thus coincide with increases in the market share of the incumbent platform.

We provide further intuition by visualizing the profit functions of the platforms for different parameter combinations and different harmful content shares chosen by the rival:

[26] Formally, $\tilde{h}_E$ solves $U_E^s(\tilde{h}_E, 1) = 0$ in this extension.

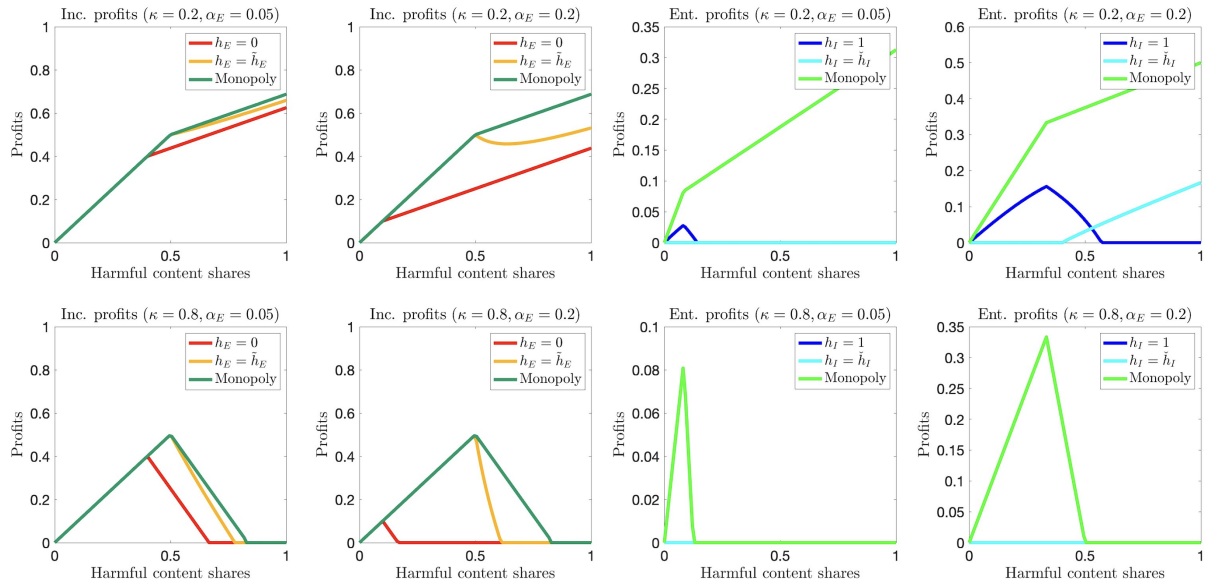

Figure C.2: Continuous users: Platform profit functions

D Details: Numerical analysis

Preliminaries:

Note that:

$$V_p(h_p) = \frac{(\theta_p + h_p(\eta_p - \theta_p))^2}{4\gamma} + (1 - h_p) - \delta h_p \implies \frac{\partial V_p(h_p)}{\partial h_p} = \frac{(\eta_p - \theta_p)(\theta_p + h_p(\eta_p - \theta_p))}{2\gamma} - (1+\delta)$$

Note further that:

$$e_p^*(h_p) = \frac{(\eta_p - \theta_p)h_p + \theta_p}{2\gamma} \implies \frac{\partial e_p^*(h_p)}{\partial h_p} = \frac{\eta_p - \theta_p}{2\gamma}$$

Calculation of relevant auxiliary functions.

We define a function $r_E(h_I)$ such that, if the incumbent sets h_I and the entrant sets h_E, rational users join the entrant if $h_E < r(h_I)$. This function solves:

$$V_E(r_E(h_I)) = V_I(h_I) \iff$$

$$\frac{(\theta_E + (\eta_E - \theta_E)r_E(\cdot))^2}{4\gamma} + 1 - (1+\delta)r_E(\cdot) = V_I(h_I) \iff$$

$$(1/(4\gamma))\big[(\theta_E)^2 + 2\theta_E(\eta_E - \theta_E)r_E(\cdot) + (\eta_E - \theta_E)^2(r_E(\cdot))^2\big] + 1 - (1+\delta)r_E(\cdot) = V_I(h_I)$$

$$\iff$$

$$\frac{(\eta_E - \theta_E)^2}{4\gamma}(r_E(\cdot))^2 + \frac{\theta_E(\eta_E - \theta_E)}{2\gamma}(r_E(\cdot)) + \frac{1}{4\gamma}(\theta_E)^2 + 1 - (1+\delta)r_E(\cdot) - V_I(h_I) = 0 \quad \text{(D.1)}$$

The solution $r_E(\cdot)$ is then given by application of the quadratic equation, with $a_E = \frac{(\eta_E - \theta_E)^2}{4\gamma} > 0$, $b_E = \frac{\theta_E(\eta_E - \theta_E)}{2\gamma} - (1+\delta) < 0$, and $c_E(h_I) = \frac{1}{4\gamma}(\theta_E)^2 + 1 - V_I(h_I) < 0$. Note that $c'(h_I) > 0$ because $\frac{\partial V_I(h_I)}{\partial h_I} < 0$. Thus, we have:

$$r_E(h_I) = \frac{-b_E - \sqrt{b_E^2 - 4a_E c_E(h_I)}}{2a_E} \quad \text{(D.2)}$$

This implies that:

$$\frac{\partial r_E(h_I)}{\partial h_I} = \frac{-2a_E(b_E^2 - 4a_E c_E(h_I))^{-0.5}\frac{\partial V_I(h_I)}{\partial h_I}}{2a_E} \quad \text{(D.3)}$$

Equivalently, we define a function $r_I(h_E)$ such that, if the entrant sets h_E and the incumbent sets h_I, all rational users join the incumbent if $h_I < r_I(h_E)$. This functions solves:

$$V_I(r_I(h_E)) = V_E(h_E) \iff$$

$$(1/(4\gamma))\big[(\theta_I)^2 + 2\theta_I(\eta_I - \theta_I)r_I(\cdot) + (\eta_I - \theta_I)^2(r_I(\cdot))^2\big] + 1 - (1+\delta)r_I(\cdot) - V_E(h_E) = 0$$

The solution $r_I(\cdot)$ is then given by application of the quadratic equation with $a_I = \frac{(\eta_I - \theta_I)^2}{4\gamma}$, $b_I = \frac{\theta_I(\eta_I - \theta_I)}{2\gamma} - (1+\delta)$, and $c_I = \frac{1}{4\gamma}(\theta_I)^2 + 1 - V_E(h_E)$. Thus, we have:

$$r_I(h_E) = \frac{-b_I - \sqrt{b_I^2 - 4a_I c_I(h_E)}}{2a_I} \quad \text{(D.4)}$$

This implies that:

$$\frac{\partial r_I(h_E)}{\partial h_E} = \frac{-2a_I(b_I^2 - 4a_I c_I(h_E))^{-0.5}\frac{\partial V_E(h_E)}{\partial h_E}}{2a_I} \quad \text{(D.5)}$$

For any h_E, the profits the entrant obtains are given by:

$$e_E^*(h_E)\rho\big[1 - F_I(r_I(h_E))\big]$$

For any such h_E, find the $h_I \in [\underline{h}_I, \tilde{h}_I]$ such that $h_I = r_I(h_E)$, i.e. where $r_E(h_I) = h_E$. Thus,

any $h_I \in [\underline{h}_I, \tilde{h}_I]$ needs to solve:

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

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:

$$F_I(h_I) = 1 - \frac{e_E^*(\underline{h}_E)}{e_E^*(r_E(h_I))} \tag{D.6}$$

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:

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

For any h_E, we thus need to have:

$$e_I^*(r_I(h_E))\big[(1-\rho) + \rho\big(1 - F_E(h_E)\big)\big] = e_I^*(\underline{h}_I) \tag{D.7}$$

Solving for $F_E(h_E)$ yields:

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

$$\iff$$

$$F_E(h_E) = \frac{1}{\rho} - \frac{e_I^*(\underline{h}_I)}{\rho e_I^*(r_I(h_E))} \tag{D.8}$$

Calculating the expected harmful content shares and market shares

We now derive expressions for the expected harmful content shares set by either platform. To see this, note that the density of harmful content shares chosen by the incumbent on $h_I \in [\underline{h}_I, \bar{h}_I)$ is given by:

$$f_I(h_I) = \frac{e_E^*(\underline{h}_E)}{[e_E^*(r_E(h_I))]^2} \frac{\partial e_E^*(r_E(h_I))}{\partial h_E} \frac{\partial r_E(h_I)}{\partial h_I}$$

Furthermore, note that the density $f_E(h_E)$ is given by the following for all $h_E \in [\underline{h}_E, \tilde{h}_E)$:

$$f_E(h_E) = \frac{e_I^*(\underline{h}_I)}{\rho[e_I^*(r_I(h_E))]^2} \frac{\partial e_I^*(r_I(h_E))}{\partial h_I} \frac{\partial r_I(h_E)}{\partial h_E}$$

The expected harmful content share chosen by the incumbent is thus:

$$\bar{H}_I = \int_{\underline{h}_I}^{\tilde{h}_I} h_I f_I(h_I) dh_I + [1 - F_I(\tilde{h}_I)](1)$$

The expected harmful content share chosen by the entrant is given by:

$$\bar{H}_E = \int_{\underline{h}_E}^{\tilde{h}_E} h_E f_E(h_E) d_E + [1 - F_E(\tilde{h}_E)](\tilde{h}_E)$$

The market share of the incumbent is:

$$\bar{M}_I = (1-\rho) + \rho \int_{\underline{h}_I}^{\tilde{h}_I} \int_{\underline{h}_E}^{\tilde{h}_E} \mathbb{1}\big[V_I(h_I) > V_E(h_E)\big] f_I(h_I) f_E(h_E) dh_I dh_E + \rho\big[1 - F_E(\tilde{h}_E)\big] F_I(\tilde{h}_I)$$

Users' expected utility:

In a pure-strategy equilibrium (h_I^*, h_E^*), the true utilities of rational and naive users are given by:

$$U^{r,*} = \mathbb{1}[j_r = I] V_I(h_I^*) + \mathbb{1}[j_r = E] V_E(h_E^*)$$

$$U^{n,*} = \mathbb{1}[j_n = I] V_I(h_I^*) + \mathbb{1}[j_n = E] V_E(h_E^*)$$

In the mixed-strategy equilibrium that emerges if the incumbent's competitive advantage is sufficiently large, naive users' expected utility is:

$$U^{p,*} = \int_{\underline{h}_I}^{\bar{h}_I} V_I(h_I) f_I(h_I) dh_I + [1 - F_I(\tilde{h}_I)] V_I(1),$$

given that naive users always visit the incumbent in equilibrium.

Rational users' expected utility is given by

$$U^{r,*} = \int_{\underline{h}_I}^{\tilde{h}_I} \int_{\underline{h}_E}^{\tilde{h}_E} \big(\mathbb{1}\big[V_I(h_I) \geq V_E(h_E)\big] V_I(h_I) + \mathbb{1}\big[V_I(h_I) < V_E(h_E)\big] V_E(h_E)\big) f_I(h_I) f_E(h_E) dh_I dh_E$$

$$+[1-F_E(\tilde{h}_E)]\int_{\underline{h}_I}^{\tilde{h}_I} V_I(h_I)f_I(h_I)dh_I+[1-F_I(\tilde{h}_I)]\left[[1-F_E(\tilde{h}_E)]V_E(\tilde{h}_E)+\int_{\underline{h}_E}^{\tilde{h}_E} V_E(h_E)f_E(h_E)dh_E\right]$$

E Proofs: Extensions

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:

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

$$\iff$$

$$\eta_I h_I + \theta_I(1-h_I) = \eta_E h_E + \theta_E(1-h_E) \tag{E.2}$$

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 (E.1) 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 incumbent 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, 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.[27] 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.

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 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 $V_p(0, s_p) > 0$ 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 join the incumbent but attain utility

[27] 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.

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 10:

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 weakly 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.[28] 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 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 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 Lemma 7:

Part 1: There exists no pure-strategy equilibrium in which some users visit the entrant.

Suppose, for a contradiction, that there exists a pure-strategy equilibrium in which some users visit the entrant and all users are indifferent between the two platforms (i.e., $V_E(h_E^*) =$

[28] 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.

$V_I(h_I^*)$ and $V_E^n(h_E^*) = V_I^n(h_I^*)$ hold). Then, $h_E^* < h_I^*$ must hold, which implies that $h_I^* > 0$. Since all users are indifferent between visiting either platform, one platform would have incentives to deviate by slightly reducing the share of harmful content it displays, since all users will join it after the deviation (note that this argument relies on the fact that we are considering equilibria in which some users visit the entrant). This is a contradiction.

Suppose, for a contradiction, that there exists an equilibrium in which some users visit the entrant and $V_E(h_E^*) = V_I(h_I^*)$ holds, i.e., rational users are indifferent between joining either platform in equilibrium. Then, $h_E^* < h_I^*$ must hold. All rational users must visit the incumbent in equilibrium—else, the incumbent would prefer to deviate from the equilibrium by marginally reducing h_I. By the previous logic, naive users cannot be indifferent. Given that the entrant must be visited by some users, naive users visit the entrant in equilibrium (and must strictly prefer to do so, since they cannot be indifferent). Thus, $V_E^n(h_E^*) = 0$ must hold. Else, the entrant could marginally increase h_E without reducing its demand. Since $\alpha < 1$, this implies that $V_E(h_E^*) < 0$ holds (since $\alpha h_E^* < h_E^*$ and $V_E(h_E)$ is strictly decreasing in h_E). Thus, rational users would obtain strictly negative utility by visiting the entrant. But thus, they cannot be indifferent in equilibrium (since the incumbent would then not be visited by any users), a contradiction.

Suppose, for a contradiction, that there exists an equilibrium in which some users visit the entrant and in which $V_E(h_E^*) \neq V_I(h_I^*)$. Suppose rational users visit the entrant. Since rational users cannot be indifferent, $V_E(h_E^*) = 0$ must hold. Thus, $V_E^n(h_E^*) > 0$ and $h_E^* > 0$ must hold. Hence, naive users must obtain strictly positive perceived utility in equilibrium. Some naive users must visit the incumbent—if no users visit the incumbent, the incumbent would deviate. Naive users cannot strictly prefer to visit the incumbent in equilibrium—else, the incumbent would slightly increase the share of harmful content it displays. Thus, naive users must be indifferent. But then, the entrant (who is not visited by all naive users in equilibrium) would strictly prefer to slightly reduce the harmful content share it displays to attract all naive users (the deviation exists because $h_E^* > 0$), a contradiction. Similar arguments establish that there cannot be an equivalent equilibrium where rational users visit the incumbent.

Part 2: In a PSE in which all users visit the incumbent, $h_E^* = 0$ and $h_I^* = \check{h}_I$ must hold.

Consider such an equilibrium. Note firstly that rational users must obtain strictly positive utility when visiting the incumbent—else, the entrant would prefer to deviate by setting $h_E = 0$, which attracts all rational users. This implies that naive users must also obtain

strictly positive perceived utility by visiting the incumbent. In the following, we say that the incentive constraint of a consumer type is slack if this consumer type strictly prefers to visit the incumbent, and that it binds if this consumer type is indifferent between visiting either platform. Note that the participation constraint of both user types must be slack (i.e., that both types obtain strictly positive utility by visiting the incumbent) by previous arguments.

In equilibrium, exactly one incentive constraint must bind, while the other incentive constraint must be slack. Suppose, for a contradiction, that both incentive constraints bind. Then, $h_I^* > 0$ and $h_E^* > 0$ must hold (if $h_E^* = 0$, naive and rational users obtain the same perceived utility by visiting the entrant, but not when visiting the incumbent—this, both types cannot be indifferent simultaneously). But then, the entrant would strictly prefer to slightly reduce the share of harmful content it displays to attract all users, a contradiction. Suppose, for a contradiction, that both incentive constraints are slack. Then, the incumbent would prefer to deviate by slightly increasing the share of harmful content it displays (because both participation constraints must be slack, by previous arguments).

In equilibrium, $h_E^* = 0$ must hold. Suppose, for a contradiction, that $h_E^* > 0$. Since the entrant is not visited by any users, it prefers to deviate by slightly reducing its harmful content share to attract users (note that at least one incentive constraint must bind in equilibrium), a contradiction.

In equilibrium, the incentive constraint of rational users must bind. Suppose, for a contradiction, that the incentive constraint of naive users binds, i.e., that $V_E^n(0) = V_I^n(h_I^*)$, while the incentive constraint of rational users is slack. Note that $V_E^n(0) = V_E(0)$ because $\alpha h = h$ if $h = 0$. Note further that $V_I(h_I^*) < V_I^n(h_I^*)$ because $h_I^* > 0$. Taken together, these inequalities imply that

$$V_E(0) = V_E^n(0) = V_I^n(h_I^*) > V_I(h_I^*), \tag{E.3}$$

which implies that rational users would strictly prefer to visit the entrant, a contradiction.

Thus, the incentive constraint of rational users must bind. Hence, $h_I^* = \check{h}_I$ must hold. Moreover, all users must visit the incumbent in equilibrium by previous arguments.

Part 3: Equilibrium existence conditions.

Consider the pure-strategy equilibrium in which $h_E^* = 0$ and $h_I^* = \check{h}_I$. By construction, $V_E^n(h_E^*) = V_E(h_E^*) = V_I(h_I^*)$. Because $V_I(h_I^*) < V_I^n(h_I^*)$ holds for any $h_I^* > 0$, all users thus prefer to visit the incumbent.

The entrant has no profitable deviations. It cannot attract any users, since any deviation

above $h_E^* = 0$ will reduce the utility of rational users and the perceived utility of naive users.

The most profitable deviation for the incumbent is to set $h_I = h_I' > \check{h}_I$, where h_I' solves $V_I^n(h_I') = V_E^n(0)$. Reductions of h_I are not profitable, since they reduce engagement but leave demand unaffected. Increases of h_I induce rational users to leave the incumbent. Thus, the most profitable deviation for the incumbent is to set $h_I = h_I'$. When deviating this way, the incumbent is visited by all naive users while maximizing engagement by these users. This deviation is not profitable if and only if $(1-\rho)e_I^*(h_I') \leq e_I^*(\check{h}_I)$. ■

Proof of Proposition 11:

Part 1: In any equilibrium in which all users visit the incumbent with probability 1, all users obtain the utility $V_I(\check{h}_I)$.

In the the pure-strategy equilibrium in which all users visit the incumbent, this follows directly because $h_I^* = \check{h}_I$ holds true in this equilibrium.

Now consider any mixed-strategy equilibrium in which all users visit the incumbent with probability 1. In equilibrium, the incumbent must set $\check{h}_I$ with probability 1. It would never be optimal to set a harmful content share below $\check{h}_I$. If the incumbent sets a harmful content share above $\check{h}_I$ with positive probability, the entrant would strictly prefer to deviate by setting $h_E = 0$, since this attracts some rational users. By implication, all users thus obtain the utility $V_I(\check{h}_I)$.

Part 2: In any equilibrium in which the entrant is visited by some users, all users either obtain the utility $U_I(\check{h}_I)$ or utility strictly below this.

By previous arguments, there exists no PSE in which the entrant is visited by some users. Thus, consider an MSE in which the entrant is visited by some users. In such an equilibrium, the incumbent must set $\check{h}_I$ with probability below 1. Suppose, for a contradiction, that there exists a MSE in which the entrant is visited by a positive measure of users and the incumbent sets $\check{h}_I$ with probability 1. Then, the entrant would only be visited if it sets $h_E = 0$. The entrant would obtain zero profits when setting any other harmful content share. Thus, the entrant can never obtain the same profits by setting $h_E = 0$ and any $h_E > 0$. Hence, we cannot have a mixed-strategy equilibrium with the stated property, a contradiction.

This implies the result. In any MSE in which the entrant is visited by some users, the incumbent sets $\check{h}_I$ with probability below 1. All users who visit the entrant obtain utility

weakly below $V_E(0) = V_I(\check{h}_I)$. All users who visit the incumbent obtain utility weakly below $V_I(\check{h}_I)$. Since the incumbent must obtain demand when setting $h_I > \check{h}_I$, some users thus visit the incumbent and obtain utility strictly below $V_I(\check{h}_I)$. ■

Proof of Proposition 12:

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 incumbent.

Part 2: There exists a ρ^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 ρ^1 such that $(1 - \rho^1)e_I^*(1) = e_I^*(\tilde{h}_I)$. If $\rho < \rho^1$, then $(1 - \rho)e_I^*(1) > e_I^*(\tilde{h}_I)$. 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 ρ^2 s.t., if $\rho > \rho^2$, there is a unique equilibrium in which $h_I^* = \check{h}_I$ and all users join the incumbent.

Now, we define ρ^2 such that $(1 - \rho^2)e_I^*(1) = e_I^*(\check{h}_I)$ If $\rho > \rho^2$, then $(1 - \rho)e_I^*(1) < e_I^*(\check{h}_I)$. 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 $V_E(h_E) \leq V_I(\check{h}_I)$ 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 Γ_I and Γ_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, $V_E(\bar{h}_E) = V_I(\bar{h}_I)$ 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 $h_I = \bar{h}_I$.[29] Thus, Γ_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 Γ_E cannot have an atom at $\bar{h}_E$ (else, both platforms would prefer to set a harmful content share 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. ■

Proof of Proposition 13:

Part 1: If $h_E = 0$ and $h_I = \check{h}_I$, all users prefer to visit the incumbent.

The value $\check{h}_I$ solves $U_I^s(\check{h}_I, 1) = U_E^s(0,1)$. If $h_I = \check{h}_I$ and $h_E = 0$, we thus have that:

$$b_I(\check{h}_I) - c(\check{h}_I) = b_E(0) - \underbrace{c(0)}_{=0} \iff U_I^s(\check{h}_I, 1) = U_E^s(0,1)$$

This implies that $b_I(\check{h}_I) - tc(\check{h}_I) > b_E(0)$ holds for every $t < 1$, which implies the desired result.

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

Part 2: There are three groups of consumers (identified by their type t_i), labeled "L", "M", and "H".

- For consumers in group L, $\frac{\partial U_I^s(h_I,t_i)}{\partial h_I} \geq 0$ and $\frac{\partial U_E^s(h_E,t_i)}{\partial h_E} \geq 0$ holds.
- For consumers in group M, $\frac{\partial U_I^s(h_I,t_i)}{\partial h_I} \geq 0$ but $\frac{\partial U_E^s(h_E,t_i)}{\partial h_E} < 0$ holds.
- For consumers in group H, $\frac{\partial U_I^s(h_I,t_i)}{\partial h_I} < 0$ and $\frac{\partial U_E^s(h_E,t_i)}{\partial h_E} < 0$ holds.

By our assumption, no other consumer group can exist.

Part 3: In any equilibrium in which all users visit the incumbent with probability 1, the incumbent must set $\check{h}_I$ with probability 1.

Suppose all users visit the incumbent with probability 1, but the incumbent sets $\check{h}_I$ with probability strictly below 1. If the entrant sets $h_E = 0$ and the incumbent sets $h_I > \check{h}_I$, all users with t_i in an open interval below 1 will strictly prefer to visit the entrant. This is because a user with type $t = 1$ strictly prefers to visit the entrant if $h_E = 0$ and $h_I > \check{h}_I$ by definition. Thus, the entrant would strictly prefer to deviate from the proposed equilibrium by setting $h_E = 0$, since this guarantees positive demand for the entrant (there is a positive measure of users with a type arbitrarily close to one by our assumption that $1 \in \text{supp}H$).

Suppose alternatively that the incumbent sets a $h_I' < \check{h}_I$. Then, the incumbent would strictly prefer to deviate from the equilibrium by setting $\check{h}_I$ because this would raise engagement and weakly increase demand. To see why the latter holds true, note that users in groups L or M would be attracted by the deviation. Users in group H strictly prefer to visit the incumbent if $h_I = \check{h}_I$, no matter what h_E the entrant chooses. This is because $U_E^s(\cdot)$ is decreasing in h_E, and the consumer prefers to visit the incumbent if $h_I = \check{h}_I$ and $h_E = 0$.

In any equilibrium in which all users visit the incumbent, the incumbent must thus set $h_I = \check{h}_I$ with probability 1.

Part 4: In any equilibrium in which some users visit the entrant with positive probability, the incumbent must set a $h_I > \check{h}_I$ with probability strictly above 0.

Suppose, for a contradiction, that the incumbent chooses a $h_I < \check{h}_I$ with probability strictly above zero. Then, the incumbent would prefer to deviate by setting $\check{h}_I$—see the previous arguments.

Suppose, for a contradiction, that the incumbent chooses $\check{h}_I$ with probability 1 in such an equilibrium. Then, all users in group M and H will visit the incumbent in equilibrium, as $U_E^s(h_E, t_i)$ is strictly decreasing in h_E. By implication, the entrant can only be visited by some users in group L. Thus, the entrant will set $h_E^* = 1$.

This implies that all users in group M and H will strictly prefer to visit the incumbent in equilibrium. To see why, note that they prefer to visit the incumbent if $h_E = 0$, and that their utility of visiting the entrant is strictly decreasing in h_E.

Thus, the incumbent would strictly prefer to deviate by slightly increasing its h_I. This is because the deviation would leave demand from users in groups M and H unaffected, and attract more users from group L. ■