NBER WORKING PAPER SERIES

ALIGNING PLATFORM AND USER INCENTIVES IN SOCIAL MEDIA AGE RESTRICTIONS

Leonardo Bursztyn
Rafael Jiménez-Durán
Aaron Leonard
Filip Milojević
Christopher Roth

Working Paper 35870
http://www.nber.org/papers/w35870

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
October 2026

Bursztyn is founder and CEO of NOMO Technologies, Inc., a company that develops digital wellness tools, and holds equity in the company. NOMO had no role in the design, conduct, funding, or analysis of this research. Funded by the Deutsche Forschungsgemeinschaft (DFG, German Research Foundation) under Germany's Excellence Strategy – EXC 2126/2 – 390838866. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

At least one co-author has disclosed additional relationships of potential relevance for this research. Further information is available online at http://www.nber.org/papers/w35870

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

© 2026 by Leonardo Bursztyn, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, and Christopher Roth. All rights reserved. Short sections of text, not to exceed two paragraphs, may be quoted without explicit permission provided that full credit, including © notice, is given to the source.

Aligning Platform and User Incentives in Social Media Age Restrictions
Leonardo Bursztyn, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, and Christopher Roth
NBER Working Paper No. 35870
October 2026
JEL No. D18, D62, L51, L86

ABSTRACT

Several countries are adopting social media bans for teenagers, yet early evidence suggests limited adherence to the age-based restrictions. Such bans ask platforms to remove profitable customers and teens to give up access to their online social networks. We study an alternative regulatory regime in which platforms remain liable for teens on the adult tier but may offer a safe tier that removes restricted features, such as algorithmic feeds, while preserving interactions with their relevant network. Provision of the safe tier and enforcement are complementary: teens can adhere without giving up access to their network, and platforms can retain displaced adult-tier users as lower-margin customers. This policy design makes the teen decision to avoid the adult tier less dependent on peer dynamics and increases teen welfare for a relevant range of regulatory penalties.

Leonardo Bursztyn
The University of Chicago
Department of Economics
and NBER
bursztyn@uchicago.edu

Rafael Jiménez-Durán
Bocconi University
rafael.jimenez@unibocconi.it

Aaron Leonard
The University of Chicago
aaronleonard@uchicago.edu

Filip Milojević
Massachusetts Institute of Technology
fmiloje@mit.edu

Christopher Roth
University of Cologne
Department of Economics
cproth89@gmail.com

1 Introduction

Several countries are considering restrictions on teenagers' social media use. Australia was the first to prohibit children under 16 from using several social media platforms. Early evidence suggests that the ban has changed behavior little, with estimated adherence rates as low as 5% (eSafety Commissioner, 2026c; Bursztyn et al., 2026a).

One possible explanation for the low adherence is an incentive problem for both platforms and teens. Platforms profit from underage users but incur the cost of identifying and removing their accounts, while neither teens nor their parents face legal sanctions for not adhering to the age restrictions. Teens also face a private cost of adhering: doing so can mean losing access to their online network of friends (Bursztyn et al., 2026a). These incentives can limit the extent to which platforms enforce the age restrictions and whether teens adhere to them.

We study an alternative regulatory regime that changes the product teenagers may legally use instead of excluding them from the platform.[1] Under a *safe-tier regime*, teens may use a version of the platform that restricts particular features, for example by removing the algorithmic feed or imposing time limits, while preserving access to their relevant network. The platform remains liable for each teen who uses the ordinary, adult tier.

Two requirements are important for the safe tier to achieve its intended outcomes. First, teens on the safe tier must be able to communicate with their relevant network, such as their friends and family. Otherwise, switching tiers would include much of the same social cost as leaving the platform. Second, the regulator must define and audit the restrictions required for the safe tier; otherwise, the platform may try to make the safe tier as similar as possible to the more profitable adult tier.[2]

The main idea is that the safe tier separates the decision of adhering to the age restriction from having access to the social network. On the user side, teens can adhere to the restriction while remaining connected to their friends, so enforcement only requires them to trade off the additional value of the adult-tier features against the cost of evading the restriction. On the platform side, a teen displaced from the adult tier can still remain as a lower-margin safe-tier customer—for example, if the safe tier is less engaging or if it entails time limits—rather than being lost entirely.

In our model, we allow for a behavioral bias that leads teens to overvalue features, such as the algorithmic feed, that could reduce their well-being—for example, because of the fear of missing out (Bursztyn et al., 2025), addiction, or self-control problems (Allcott et al., 2022).[3]

[1] Related policy work in child and adolescent psychiatry advocates age-tiered access to social media as a way to balance protection with continued digital participation (Köhler-Dauner et al., 2025).

[2] Recent experience with Meta's non-profiled feeds illustrates this problem. Meta allowed users to choose a feed without personalized recommendations, but the users' choice was not always respected: after closing the app or navigating elsewhere, the algorithm would go back to the profiled feed and users had to select the alternative again. A Dutch court required Meta to make the non-profiled feed easier to access and to preserve users' choice (Rechtbank Amsterdam, 2025).

[3] This is modeled through a behavioral wedge, w, between the value teens perceive from the restricted features,

A teen may therefore choose the adult tier even when they would be better off without the restricted features. We assume that the safe tier removes the features responsible for this behavioral distortion while preserving access to their relevant social network.

Another important ingredient is that the regulator cannot directly contract on the platform's effective enforcement effort. A regulator can require age-assurance procedures, such as identity or age verification, but the platform typically retains discretion over how these procedures are implemented.[4] We capture this imperfect contractibility by letting the regulator set an expected penalty for each underage user on the adult tier, while the platform chooses how strongly to enforce the restriction. The regulator also specifies the requirements for the safe tier, and the platform chooses whether to offer it.

Our model delivers three main results. First, the safe tier changes the platform's incentives to provide the product and enforce the age restriction. As a benchmark, consider an age-based ban such as the Australian one. Removing an underage user costs the platform the profit generated by that user and also requires costly screening. The platform therefore does not enforce unless the expected penalty per underage user is sufficiently high—above the expected profit per user.[5] Below this threshold, the platform does not enforce the age restriction and the equilibrium is unchanged relative to a world without a ban.

The model also explains why the platform does not offer the safe tier voluntarily (both under the ban and in the absence of any regulation). The safe tier attracts teens who would otherwise stay off the platform, generating new revenue, but it also moves some existing users away from the more profitable adult tier. If this loss from "cannibalization" is higher than the revenue from new users, the platform does not find it profitable to offer the safe tier. Regulation changes these incentives in two steps. First, a lower regulatory penalty is required to induce the platform to begin offering the safe tier, because providing it causes some teens to leave the adult tier and therefore reduces platform liability without requiring active enforcement. Second, at a higher penalty, the platform begins enforcing the adult-tier restriction because some teens

such as the algorithmic feed, and the value they truly derive from them. Thus perceived utility on the adult tier exceeds true utility by w, while perceived and true utility coincide on the safe tier.

[4] Australia's social media minimum-age policy provides an example. The law requires platforms to take "reasonable steps" to prevent under-16s from holding accounts but does not prescribe how platforms must satisfy this obligation.

[5] Such an expected penalty may be unrealistic in practice. In the Australian case, a back-of-the-envelope calculation suggests annual revenue of approximately A$155 per platform user (https://www.accc.gov.au/system/files/Digital%20platforms%20services%20inquiry%20-%20Interim%20report%206%20-%20Report%20on%20social%20media%20services_0.pdf). Applying Meta's 2022 Family of Apps operating margin of 37% gives approximately A$58 in annual profit per user (https://investor.atmeta.com/investor-news/press-release-details/2023/Meta-Reports-Fourth-Quarter-and-Full-Year-2022-Results/default.aspx). The new maximum penalty of A$109.2 million (https://minister.infrastructure.gov.au/wells/media-release/australia-doubles-down-social-media-law), spread across Instagram's reported 351,135 Australian users aged 13–15, amounts to roughly A$311 per user for one maximum fine (https://www.esafety.gov.au/sites/default/files/2025-02/Behind-the-screen-transparency-report-Feb2025.pdf). If there is a 10% probability of one maximum fine per year, this would amount to approximately A$31 per user annually, below the A$58 profit benchmark.

displaced from the adult tier can now remain on the platform as safe-tier users. If the additional safe-tier revenue generated by enforcement exceeds its cost, both responses occur at penalties for which the ban remains unenforced. Liability that is too small to induce enforcement under the ban can therefore first induce the platform to provide the safe-tier provision and then to enforce the age restriction.

Second, the safe tier changes *which* teens are affected by enforcement. Because both tiers provide access to the same relevant network, the choice between them depends on the perceived value of the incremental features available only on the adult tier and on the cost of evading the restriction. As enforcement increases, teens switch from the adult tier to the safe tier in increasing order of how much they truly value the adult feed. When the cost of accessing the adult tier exactly cancels the behavioral bias, teens harmed by the adult feed choose the safe tier, while those who benefit from it remain on the adult tier. Under the ban, by contrast, a teen's adherence to the age restriction depends on their perceived full value of the platform, including the network component. The safe tier therefore targets those teens most harmed by the adult feed while keeping their social connections, whereas the ban excludes teens who least value the platform, who need not be those most harmed by it.

Third, maximum teen welfare is higher with the safe tier than under either the ban or no regulation, focusing on the relevant range of penalties for the regulator (i.e., those for which the ban remains unenforced). This comparison follows from the same cannibalization condition that explains why the platform does not offer the safe tier voluntarily. Throughout this range of penalties, the ban leaves platform behavior unchanged, whereas the regulator can induce the platform to provide the safe-tier without active enforcement. Provision alone raises welfare relative to the ban by expanding access to the network and allowing teens to avoid a harmful feed. Additionally, there could be further welfare gains from enforcement depending on how the benefits of correcting biased choices compare to the circumvention costs incurred by remaining adult-tier users and the losses from a smaller network.

We extend our baseline model in three ways. First, we relax the assumption that the safe tier achieves perfect interoperability—that is, that access to the relevant network is the same as when using the adult tier. The main provision, enforcement, and welfare results continue to hold with sufficiently small interoperability frictions. Second, we relax the assumption that the behavioral bias is homogeneous across users. Heterogeneity in the behavioral bias leaves the platform-side provision and enforcement results unchanged. When this heterogeneity is bounded, enforcement continues to sort teens according to whether the restricted feed benefits or harms them, except for teens whose true feed value is close to zero; under an additional condition linking higher perceived feed value to lower behavioral bias, the welfare results also extend. Third, we consider a reduced-form retention channel through which competition between platforms can affect enforcement. If the incumbent platform retains less of the safe-tier revenue generated by enforcement, its enforcement threshold rises. This captures one way in which

competition with lax rivals can weaken enforcement incentives.

Recent policy developments illustrate the incentive problem highlighted by our model. When adherence requires giving up access to the social network, teens have strong incentives to evade the restriction, while platforms bear the cost of enforcing it. In Australia, where under-16s were banned from holding accounts on major social media platforms in December 2025, platforms reported restricting access to approximately 4.7 million accounts after the age restriction took effect, yet changes in actual social media use were much smaller: three months later, daily use among 10–15-year-olds had changed little (eSafety Commissioner, 2026b,a).[6] Indeed, surveying Australian teenagers four months in, we found that fewer than one in three 14–15-year-olds were adhering to the ban (Bursztyn et al., 2026a). Social forces played an important role: among teens still using social media, 52% cited at least one social motive for continued use—most often that their friends still used the platforms or a fear of missing out. Teenagers reported requiring two out of three peers to stop using social media before stopping themselves, while believing that only 30% had done so, and, two months later, adherence had fallen to about one in twenty. Taken together, this evidence points to a need for regulatory designs that better align platform and user incentives. The safe tier does so on both sides: teens keep access to the relevant network, and a teen removed from the adult tier can be retained as a lower-margin safe-tier customer rather than lost entirely.

Regulators are beginning to move in this direction by going beyond binary age-based exclusion toward age-specific restrictions on the product available to teenagers. In August 2026, Meta agreed to pay up to $18 billion to settle litigation brought by U.S. states and, as part of the settlement, to impose new restrictions on teenage use of Facebook and Instagram, including limits on daily use and overnight access (Bobrowsky, 2026). Such restrictions can be understood as examples of the safe-tier approach we study: teenagers keep access to the platform and their relevant network but face restrictions on particular features or dimensions of use.

Our paper contributes most directly to the literature on platform regulation and delegated enforcement. Seminal work studies when regulators should rely on intermediaries that can detect or prevent misconduct (Kraakman, 1986). More recent work examines how liability for third-party harms affects platforms' incentives to monitor, audit, exclude users, and adjust prices and participation (Lefouili and Madio, 2022; Hua and Spier, 2025). Related work studies how regulation affects platform design when the objectives of the regulator and the platform are different (Acemoglu et al., 2024). We show that allowing a platform to retain restricted users on a different version of its product can induce it to enforce the restriction.

The paper also contributes to behavioral public economics and the literature on regulating internalities (O'Donoghue and Rabin, 2006; Allcott et al., 2014, 2019). In our model, teens perceive restricted features, such as the algorithmic feed, as more valuable than they are to

[6] Since the law imposes a platform-level obligation rather than a per-user penalty, this evidence speaks to the model's motivating forces rather than a precise test.

their true welfare. This assumption is motivated in part by evidence of habit formation and self-control problems in social-media use (Allcott et al., 2022), as well as fear of missing out (Bursztyn et al., 2025). Unlike a standard corrective tax, however, the regulator does not directly change the incentive faced by the consumer. Instead, in our setting, the regulator imposes liability on the platform, which in turn chooses what product to offer and how much to enforce. The effect on teens therefore depends on the platform's response and on whether adherence requires giving up access to their friends.

Finally, our analysis relates to the literatures on product differentiation and on compatibility in markets with network effects. A monopolist may withhold a lower-quality product when it cannibalizes demand for a more profitable version (Mussa and Rosen, 1978; Deneckere and McAfee, 1996). The same force can lead the platform in our model to withhold the safe tier, while regulatory liability can reverse this incentive. The safe tier also differs from a standard lower-quality product: removing a feature that consumers value can nevertheless increase true utility for teens who are harmed by that feature. Because the value of the safe tier depends on continued access to the platform's network, our analysis also builds on the literature on compatibility and network effects (Katz and Shapiro, 1985; Crémer et al., 2000; Farrell and Saloner, 1985; Rosaia, 2025; Bursztyn et al., 2026b).

2 Model

In this section, we study how introducing a safe tier changes the incentives of both teens and a social media platform under age-based restrictions. We compare a ban, under which teens are prohibited from using the platform, with a safe-tier regime, under which the platform may legally serve teens a version of the product that preserves communication with their social network but removes potentially harmful features, such as the algorithmic feed. The key mechanism is that the safe tier separates adhering with the age restriction from accessing the social network. On the user side, with an interoperable safe tier, a teen can adhere to the age restriction while remaining connected to their friends. Enforcement only requires the teen to trade off the additional value of the adult-tier features against the cost of evading the restriction. On the platform side, the safe tier allows the platform to retain a teen displaced from the adult tier as a lower-margin customer rather than losing them entirely. These two forces make the safe tier and enforcement efforts complementary in terms of reducing teen demand for the adult tier. This complementarity implies that introducing a safe tier can induce the platform to provide a compliant product and enforce age restrictions, where a ban would leave platform behavior unchanged.

2.1 Setup and Unregulated Benchmark

Teens. There is a unit mass of teens. Depending on the regulatory regime and the platform's product offering, each teen chooses among an adult tier A, a safe tier S, and staying off the platform. The safe tier removes the adult-tier feed but preserves communication with peers: in the baseline model, the two tiers are perfectly interoperable, so teens can interact with the same social network on either tier.

Teens have both perceived and true utility from using both tiers. A teen's perceived utility is:

$$U^S(q) = u(q) + \varepsilon,$$
$$U^A(q, c) = u(q) + \varepsilon + w + \theta - c,$$

where q is the size of the teen network across both tiers and u is continuously differentiable and strictly increasing. We focus on the endogenous network of teenagers q and abstract from adults and advertisers in our model. A key aspect of the mechanism is how much of the relevant network a teen loses by complying, and related, to what extent safe-tier users can interact with adult-tier users. We relax the perfect interoperability assumption in Appendix B.1.

The term ε, drawn from distribution F, is the teen's membership benefit: the idiosyncratic value of access to the platform.

The parameter $w > 0$ is a behavioral bias wedge. Due, for example, to addiction, fear of missing out, biased beliefs, or self-control problems, teens overvalue the adult-tier feed by w. Note that teens do not misperceive the true utility from the safe tier. For simplicity, we assume that the bias is constant across teens, but we relax this assumption in Appendix B.2.

The term θ, drawn from distribution G, is the true additional value that the teen derives from the adult-tier feed relative to the safe tier.[7] Thus, $\theta < 0$ means that the adult feed lowers the teen's true utility, whereas $\theta > 0$ means that the teen benefits from it. Teens with $\theta < -w$ are therefore harmed by the feed and perceive it as harmful; teens with $-w < \theta < 0$ are harmed by the feed but perceive it as beneficial; and teens with $\theta > 0$ both benefit from the feed and perceive it as beneficial.

The term $c \geq 0$ is a utility cost set by the platform through its enforcement technology and borne by a teen who accesses the restricted adult tier. It captures, for example, circumvention effort, borrowing credentials, and verification frictions.

We normalize both perceived and true utility from not using the platform to zero. For simplicity, we ignore non-user externalities as in Bursztyn et al. (2025), and capture patterns such as fear of missing out with the reduced-form term w. Our welfare criterion is teen welfare,

[7] We assume that F and G have continuously differentiable densities f and g, with f, g, f', and g' bounded, that ε and θ are independent and have finite means, and that ε has full support. We assume $0 < G(-w) < 1$, so that in the absence of enforcement there is a positive mass of teens who prefer each tier.

defined as aggregate true utility.

Platform and regulation. The platform earns a margin $\pi > 0$ per adult-tier user and $\delta\pi$ per safe-tier user, where $\delta \in (0, 1)$ captures that the safe tier is harder to monetize (e.g., engagement may be lower without the algorithmic feed). The platform chooses whether to offer the safe tier when legally permitted and chooses a policy to enforce the age restriction. This policy is captured by the disutility term c which entails cost $C(c)$ for the platform. We assume C is twice continuously differentiable, strictly increasing, and convex, with $C(0) = 0$.

A teen using the adult tier exposes the platform to expected regulatory liability $\tau \geq 0$, while safe-tier users do not entail any liability (depending on the regulatory regime). The regulator sets τ and the requirements for a safe tier, but cannot directly contract on enforcement effort c. Therefore, the platform decides whether to offer the safe tier and how strongly to enforce the adult-tier restriction.

Assumption 1 (Limited network feedback)**.**

$$\sup_x f(x) \cdot \max_{q \in [0,1]} u'(q) < 1.$$

Assumption 1 rules out multiple participation equilibria driven purely by network effects. As the network grows, the platform becomes more valuable and additional marginal teens may join. The strength of this feedback depends on the network effect u' and the density of marginal teens, f. This assumption allows us to abstract from equilibrium multiplicity and tipping forces (Bursztyn et al., 2026a) and focus on the platform's incentives to provide the safe tier and enforce the age restriction.

Unregulated benchmark. We begin by analyzing the platform's incentives without regulation, i.e., when $\tau = 0$. First, suppose that the platform only offers the adult tier. Let H denote the CDF of the perceived standalone value of adult-tier access, $\varepsilon + w + \theta$. A teen participates whenever $u(q) + \varepsilon + w + \theta \geq 0$, so equilibrium use q_0 solves

$$q_0 = 1 - H\big(-u(q_0)\big).$$

Therefore, marginal participants have zero perceived surplus but a true negative private surplus of $-w$. Their participation also creates a network benefit $q_0 u'(q_0)$ for existing users, so participation is socially excessive at the margin whenever $q_0 u'(q_0) < w$.

To see whether the unregulated platform would itself provide a safe tier, suppose instead

that it offers both tiers at $c = 0$. Equilibrium participation $\tilde{q}_0$ satisfies

$$\tilde{q}_0 = \underbrace{G(-w)\big[1 - F(-u(\tilde{q}_0))\big]}_{\tilde{q}_{S0}} + \underbrace{\int_{-w}^{\infty} \big[1 - F(-u(\tilde{q}_0) - w - \theta)\big]\, dG(\theta)}_{\tilde{q}_{A0}},$$

where $\tilde{q}_{S0}$ and $\tilde{q}_{A0}$ denote safe- and adult-tier use. Offering the safe tier expands the available choice set, so overall platform demand rises, $\tilde{q}_0 > q_0$, and a positive mass of teens chooses the safe tier, $\tilde{q}_{S0} > 0$. The change in platform profit is $\pi\,[\tilde{q}_0 - q_0 - (1-\delta)\tilde{q}_{S0}]$. The term $\tilde{q}_0 - q_0$ captures the market-expansion channel from the additional revenue from a bigger network, while $(1-\delta)\tilde{q}_{S0}$ captures the revenue lost because the platform charges a lower markup on safe-tier users.

Condition 1 (Cannibalization)**.**

$$\underbrace{\tilde{q}_0 - q_0}_{\text{market expansion}} \quad < \quad \underbrace{(1-\delta)\tilde{q}_{S0}}_{\text{cannibalization}}.$$

Condition 1 is necessary and sufficient for the unregulated platform to withhold the safe tier. It holds when the safe tier is sufficiently harder to monetize and when making it available does not expand participation enough to offset the cannibalization of adult-tier users.[8] We take this as the economically interesting benchmark: there is latent demand for a safer product, but the platform does not provide it because serving that demand requires moving users to a less profitable tier.

2.2 Ban

Equilibrium under the ban. Next, consider the case when the regulator prohibits teen use of the platform and imposes expected liability τ for each teen who uses it. The safe tier is unavailable, so a teen chooses between accessing the banned adult tier and leaving the platform. Equilibrium use $q_B(c)$ solves

$$q_B(c) = 1 - H\big(c - u(q_B(c))\big).$$

The platform solves

$$\max_{c \geq 0} \ \Pi_B(c;\tau) = (\pi - \tau) q_B(c) - C(c).^{9} \tag{1}$$

[8] A sufficient condition based only on primitives is $(1-\delta)G(-w)\big[1-F(-u(0))\big] > \frac{\sup_x f(x)}{1-\kappa} \int_{-\infty}^{-w} (-w-\theta)\, dG(\theta)$, where $\kappa \equiv \sup_x f(x) \max_{q\in[0,1]} u'(q)$. The left-hand side is a lower bound on the revenue loss from cannibalization, while the right-hand side bounds the increase in participation generated by making the safe tier available, including its amplification through network effects.

[9] We assume strict concavity of the ban enforcement problem for $\tau \geq \pi$. A sufficient condition is $q_B''(c) \geq 0$ and $C''(c) > 0$.

Proposition 1 (Ban enforcement)**.** *Under Assumption 1, define*

$$\tau_B \equiv \pi + \frac{C'(0)}{|q'_B(0)|}. \tag{2}$$

Then $\tau_B \geq \pi$. The platform chooses $c^*_B(\tau) = 0$ for $\tau \leq \tau_B$ and positive enforcement for $\tau > \tau_B$. Equilibrium enforcement is continuous in τ and strictly increasing once it becomes positive.

As long as the expected liability from an underage user is smaller than the markup that the platform can charge for them, enforcement of the age restriction reduces revenue and requires costly screening. The platform therefore cannot begin to enforce before $\tau = \pi$. Even above π, the regulatory saving from reducing adult-tier use must cover the marginal cost of enforcement, which explains the additional term in τ_B. The ban can therefore remain ineffective at causing teens to leave the platform even after the expected penalty exceeds the adult-tier markup.

This lack of platform enforcement does not imply that enforcement is underprovided from the point of view of teen welfare. A small increase in enforcement reduces the amount of teens who overuse the platform due to the bias, but it also imposes circumvention costs on those who remain and contracts the network.

2.3 Safe Tier

Teen Incentives. Now consider the case in which a regulator continues to penalize teen use of the adult tier but allows the platform to operate an unpenalized safe tier. The safe tier changes the teen's adherence problem. Under the ban, a teen who adheres to the age restriction trades off accessing the adult-tier vs. leaving the platform. The teen chooses to ignore the age restriction if $u(q) + \varepsilon + w + \theta - c > 0$. Adherence therefore requires giving up not only the adult-tier feed but also the value of the network.

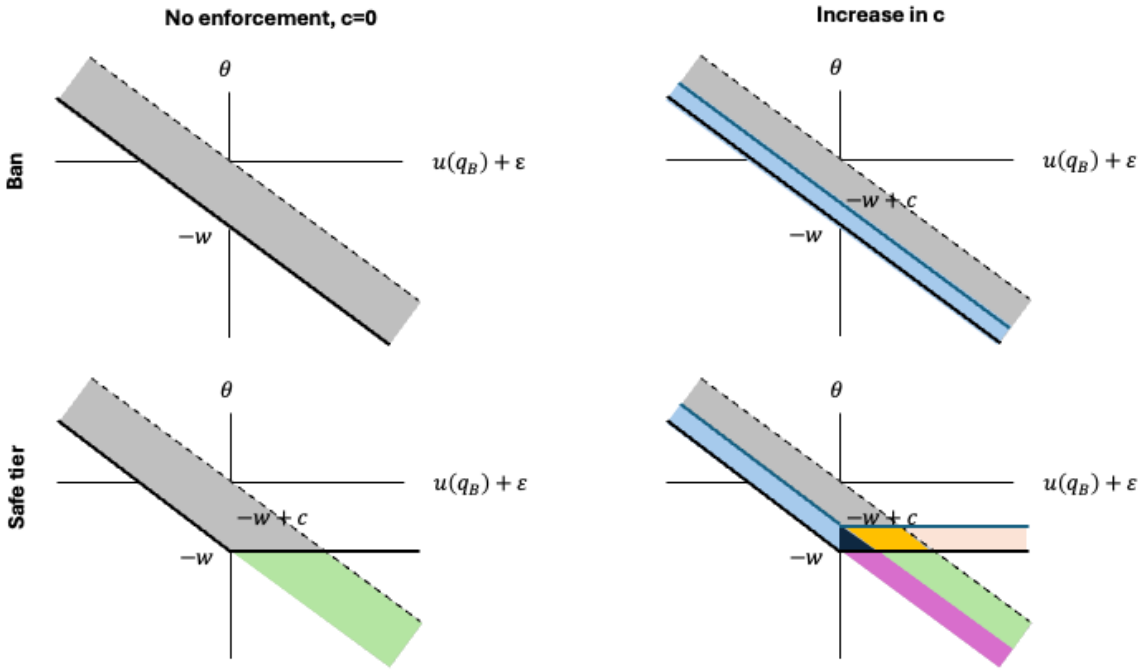
With the safe tier, a participating teen instead compares $u(q) + \varepsilon + w + \theta - c$ with $u(q) + \varepsilon$. so the membership and network terms cancel. Conditional on participating, the teen chooses the safe tier if and only if $\theta < c - w$. Thus, with perfect interoperability, enforcing the age restriction requires it to be costly enough to overcome the perceived additional value of the adult-tier features, $w + \theta$, rather than the value of the entire platform.

This change in tradeoffs results in a complementarity between the safe tier and the enforcement of the age restriction. Figure 1 provides intuition of how teen incentives to use the platform change between the case of the ban (top panels) and the case of the safe tier (bottom panels) and how these incentives respond after enforcement moves from zero (left panels) to a small c (right panels). The subfigures display the cutoffs that are used to determine which teens sort into the different tiers available either under the ban or the safe-tier scenarios. The relevant parameter space for these cutoffs is given by the true utility of using the safe tier, $u(q_0) + \varepsilon$, and the additional true utility of the adult tier vs. the safe tier, θ. For simplicity, we hold constant

the size of the network at its equilibrium value under the ban, q_B.

The top left panel illustrates the equilibrium under the ban. The solid black line represents parameter values that make teens indifferent between using the adult tier and not using the platform; teens above the line use the platform and teens under the line do not. The gray area constitutes the set of teens who use the platform despite being harmed by it. The bottom left panel illustrates that the safe tier reduces the share of harmed teens who use the adult tier (green area). The top right panel shows that an increase in enforcement under the ban would lead some teens to stop using the platform instead of using the adult tier (light blue rectangle). The bottom right panel shows that, holding the network fixed, the same increase in enforcement causes a smaller reduction in overall platform demand under the safe tier, because some teens switch tiers rather than exit. The dark blue and pink areas denote the share of teens who would exit the platform under the ban but either remain on the safe tier even in the absence of enforcement (pink) or are induced to switch from the adult to the safe tier in response to the higher enforcement (dark blue). The yellow and peach areas denote the share of teens who would remain in the adult tier under the ban but instead switch to the safe tier. The yellow rhomboid in particular is the set of harmed teens who are targeted by the increase in enforcement under the safe tier who are unaffected in the case of the ban.

Figure 1: How the Safe Tier Changes Adherence



Notes: The figure holds the teen network fixed at its ban-equilibrium value, q_B. The top row shows the ban and the bottom row the safe-tier regime; the left column sets $c = 0$ and the right column considers higher enforcement. The diagonal lines are adult-tier participation cutoffs. Under the safe tier, the horizontal cutoff $\theta = c - w$ separates adult- from safe-tier use, while the vertical cutoff determines safe-tier participation. Under a ban, stronger enforcement reduces adult-tier use only by inducing exit. With a safe tier, it can instead shift teens from the adult tier to the safe tier while preserving network access.

Platform incentives. The safe tier also changes the platform's incentive to enforce the age restriction. Holding network size fixed, moving a teen off the adult tier under a ban changes platform profit by $\tau - \pi$: the platform avoids the liability but loses the full adult-tier margin. If the teen instead moves from A to S, the corresponding gain is $\tau - (1-\delta)\pi$, since the platform retains the safe-tier margin $\delta\pi$. The safe tier therefore makes enforcement less costly to the platform.

Equilibrium participation when both tiers are offered satisfies

$$q(c) = \underbrace{G(c-w)\big[1 - F(-u(q(c)))\big]}_{q_S(c)} + \underbrace{\int_{c-w}^{\infty} \big[1 - F(c - u(q(c)) - w - \theta)\big]\, dG(\theta)}_{q_A(c)}, \tag{3}$$

where $q_S(c)$ and $q_A(c)$ denote safe- and adult-tier use. Platform profit is

$$\Pi_S(c;\tau) = \delta\pi q(c) + \big[(1-\delta)\pi - \tau\big] q_A(c) - C(c). \tag{4}$$

To isolate the decision of the platform to offer the safe tier, first hold enforcement fixed at $c = 0$. Relative to offering only the adult tier, the profit gain from opening the safe tier is

$$\Pi_S(0;\tau) - \Pi_B(0;\tau) = \underbrace{\pi\left[\tilde{q}_0 - q_0 - (1-\delta)\tilde{q}_{S0}\right]}_{\text{cannibalization and market expansion}} + \underbrace{\tau\left(q_0 - \tilde{q}_{A0}\right)}_{\text{regulatory liability avoided}}. \tag{5}$$

Condition 1 makes the first term negative. It also implies $q_0 - \tilde{q}_{A0} > \delta\tilde{q}_{S0} > 0$, so offering the safe tier reduces adult-tier use even without enforcement. Thus, the second term is positive and increases with the regulatory penalty. The next proposition shows how the penalty first induces safe-tier provision and, at a higher level, enforcement of the adult-tier restriction. For the safe tier to induce enforcement at penalties that leave the ban unenforced, the revenue retained from displaced adult-tier users must provide a sufficiently strong incentive to enforce. The following condition is sufficient for this to occur.

Condition 2 (Early enforcement)**.** There exists some $c > 0$ such that

$$\delta\pi\left[q_S(c) - q_S(0)\right] > C(c).$$

How restrictive is Condition 2? It is equivalent to $\Pi_S(c;\pi) > \Pi_S(0;\pi)$ for some $c > 0$. Since the ban cannot induce enforcement below $\tau = \pi$, the condition requires that some positive enforcement already be profitable under the safe-tier regime by that point.[10]

[10] A simple local sufficient condition is $\delta\pi q_S'(0) > C'(0)$: the marginal safe-tier revenue generated by a small increase in enforcement exceeds its marginal cost. If $C'(0) = 0$, it is enough that $q_S'(0) > 0$: a small increase in enforcement brings more teens onto the safe tier than the smaller network causes to leave it. Under the strict concavity imposed below, the local condition is also necessary for Condition 2.

Proposition 2 (Safe-tier provision and enforcement)**.** *Suppose Condition 1 and Assumption 1 hold, and the safe-tier enforcement problem is strictly concave in c for* $\tau \geq (1-\delta)\pi$*.* [11] *For penalties* $\tau \leq \tau_B$*, the platform prefers to offer the safe tier if and only if*

$$\tau > \tau_P \equiv \pi \frac{(1-\delta)\tilde{q}_{S0} - (\tilde{q}_0 - q_0)}{q_0 - \tilde{q}_{A0}}, \tag{6}$$

where $0 < \tau_P < (1-\delta)\pi$*.*

Conditional on offering the safe tier, the platform chooses zero enforcement for

$$\tau \leq \tau_E \equiv \pi - \frac{\delta\pi q_S'(0) - C'(0)}{|q_A'(0)|}. \tag{7}$$

and positive enforcement for $\tau > \tau_E$*. Enforcement is continuous and strictly increasing after it becomes positive. Moreover,* $\tau_E > (1-\delta)\pi$*, so the safe tier is provided before the platform begins to enforce the adult-tier restriction.*

Under Condition 2, $\tau_E < \pi \leq \tau_B$*. The platform therefore offers only the adult tier for* $\tau < \tau_P$*, both tiers with zero enforcement for* $\tau \in (\tau_P, \tau_E]$*, and both tiers with positive enforcement for* $\tau \in (\tau_E, \tau_B]$*. For* $\tau > \tau_B$*, both regulatory regimes may induce positive enforcement, but the relative profitability of the two product offerings is not generally ranked.*

Under Condition 2, there are therefore two critical penalty levels below τ_B. Once τ exceeds τ_P, the platform finds it profitable to offer the safe tier. At the higher threshold τ_E, it also begins to enforce the age restriction on the adult tier. Thus, as the penalty rises, the platform first changes its product offering and only later its enforcement behavior.

How much earlier does the safe-tier regime induce enforcement? Under Condition 2, $\tau_E < \pi \leq \tau_B$, so there is a range of penalties for which the platform enforces with a safe tier but not under a ban.[12] The gap $\pi - \tau_E$ increases with δ and decreases with $C'(0)$, holding demand fixed. When $C'(0) = 0$, it is proportional to the net increase in safe-tier use per teen displaced from the adult tier. This accounts for existing safe-tier users who may leave as the network shrinks. Since $\tau_E > (1-\delta)\pi$, this gap is bounded above by $\delta\pi$. A higher δ therefore strengthens the platform's incentive both to provide the safe tier and to enforce the age restriction conditional on providing it.

Corollary 1 (Targeting)**.** *When both tiers are offered, a participating teen chooses the safe tier if and only if* $\theta < c_S^*(\tau) - w$*. Hence, as equilibrium enforcement rises, teens move from the adult tier to the safe tier in increasing order of their true additional value of the adult-tier feed. If equilibrium enforcement reaches* $c_S^*(\tau) = w$*, the safe tier contains exactly the participating teens with* $\theta < 0$*.*

[11] A sufficient condition is $q_A''(c) \geq 0$ and $C''(c) > \delta\pi \max\{q''(c), 0\}$ over the relevant range.

[12] The gap between the safe-tier enforcement threshold and the adult-tier margin is $\pi - \tau_E = \frac{\delta\pi q_S'(0) - C'(0)}{|q_A'(0)|} > 0$.

When $c = 0$, only teens with $\theta < -w$ voluntarily choose the safe tier: the adult tier harms them enough to perceive the harm despite the bias. Positive enforcement then reaches teens with $-w < \theta < 0$, who are also harmed by the feed but perceive it as beneficial. At $c = w$, enforcement exactly offsets the common bias wedge and tier choice is aligned with the sign of true feed value. The exact ordering in this corollary relies on the assumption of a common behavioral wedge. With heterogeneous w, enforcement sorts teens by their perceived feed value $z = \theta + w$ rather than by θ alone. Appendix B.2 shows that the platform-side results continue to apply in that case and provides a distributional condition under which higher perceived feed value is associated with higher true feed value and lower behavioral bias. Under this condition, the targeting and single-crossing welfare results extend to heterogeneous bias.

To summarize, the safe tier is complementary with enforcement of the age restriction, both on the user side and on the platform side. For users, an increase in enforcement leads some users—who would keep using the adult tier under the ban—to switch from the adult tier to the safe tier, since they now trade off the cost of not adhering to the restriction against the additional features of the adult tier, as opposed to the full value of the network. For the platform, the safe tier reduces the private cost of enforcement because teens displaced from the adult tier can still be retained as less profitable safe-tier users rather than lost entirely. Under Condition 2, these two forces allow the regulator to induce enforcement at lower expected penalties than under the ban.

2.4 Teen Welfare and the Regulatory Penalty

We now study how the provision and enforcement responses characterized above affect teen welfare. Let W_U denote teen welfare in the unregulated equilibrium, and let $W_B(c)$ and $W_S(c)$ denote teen welfare under the ban and safe-tier regimes, respectively, at enforcement level c.

Welfare effects from the provision of the safe tier. The following result makes explicit the conflict between platform incentives and teen welfare regarding the provision of the safe tier.

Proposition 3 (Underprovision of the safe tier)**.** *Suppose Condition 1 holds. In the absence of regulation, the platform withholds the safe tier, even though offering it at zero enforcement strictly raises teen welfare,* $W_S(0) > W_U$.

For the platform, Condition 1 implies that the additional participation generated by the safe tier is too small to compensate for moving users away from the more profitable adult tier. For teens, however, offering the safe tier increases perceived surplus by expanding both their choice set and the network. It also implies that $\tilde{q}_{A0} < q_0$, so fewer teens are exposed to the behavioral bias from the adult tier. Since aggregate true welfare equals aggregate perceived surplus minus w for each adult-tier user, both forces result in higher teen welfare.

Welfare effects of enforcing the age restriction. The welfare effect of enforcement is less clear. A higher enforcement helps correct the distorted choices caused by the behavioral bias, but it also imposes circumvention costs on all teens who remain on the adult tier and can reduce the size of the network. Differentiating teen welfare under the safe-tier regime gives

$$\frac{dW_S(c)}{dc} = \underbrace{w|q'_A(c)|}_{\text{bias correction}} - \underbrace{q_A(c)}_{\text{circumvention costs}} - \underbrace{q(c)u'(q(c))|q'(c)|}_{\text{network contraction}}. \tag{8}$$

The first term is the benefit from correcting the behavioral bias; each teen displaced from the adult tier does not experience anymore the wedge w between perceived and true utility. The second term captures the additional circumvention cost incurred by teens who remain on the adult tier. The third captures the loss in network value when enforcement reduces total network size. The safe tier changes the last tradeoff because a teen displaced from the adult tier can switch to the safe tier rather than exit, allowing adult-tier use to fall without necessarily producing an equivalent contraction in the network.[13]

Equation (8) can be written as $dW_S(c)/dc = |q'_A(c)|[w - K_S(c)]$, where

$$K_S(c) \equiv \frac{q_A(c)}{|q'_A(c)|} + q(c)u'(q(c))\frac{|q'(c)|}{|q'_A(c)|}$$

is the marginal social cost per teen displaced from the adult tier. Enforcement therefore increases teen welfare when the behavioral distortion w exceeds this marginal social cost. Hence, in general, $K_S(c)$ need not be monotone, so welfare can be increasing or decreasing at different levels of enforcement.

The following additional assumption on the distribution of feed value yields a simple characterization of the welfare effects of enforcement.

Assumption 2 (Decreasing hazard of adult-tier value)**.** The distribution of θ has a strictly decreasing hazard rate for $\theta \geq -w$: $g(\theta)/[1 - G(\theta)]$ is strictly decreasing over this range.

This assumption implies that, holding the network fixed, a given increase in enforcement displaces a smaller fraction of the remaining adult-tier users as enforcement rises. It holds, for example, when adult feed values have a Pareto upper tail over the relevant range. It does not hold for normal or logistic distributions, which have increasing hazards. The restriction applies throughout the range $\theta \geq -w$, not just to large feed values, and implies that the density is decreasing over this range.

Lemma 1 (Increasing marginal cost of enforcement)**.** *Suppose Assumptions 1 and 2 hold. There exists $\bar{\kappa}_W > 0$ such that, if $\|u'\|_\infty + \|u''\|_\infty < \bar{\kappa}_W$, then $K_S(c)$ is strictly increasing for*

[13]The expression is similar under the ban: $dW_B(c)/dc = w|q'_B(c)| - q_B(c) - q_B(c)u'(q_B(c))|q'_B(c)|$. The same three forces operate, but in that case every teen displaced from the adult tier also leaves the network.

$c \in [0, c_S^*(\tau_B)]$.[14]

Since $dW_S(c)/dc = |q'_A(c)|[w - K_S(c)]$, the lemma implies that teen welfare is either monotone or single-peaked over $c \in [0, c_S^*(\tau_B)]$, depending on the behavioral bias relative to marginal social cost. If $w \leq K_S(0)$, any positive enforcement reduces teen welfare. If $w > K_S(0)$, welfare initially increases. If there exists $c^W \in (0, c_S^*(\tau_B)]$ such that $K_S(c^W) = w$, this level is unique and welfare decreases beyond it; otherwise, welfare continues to increase.

Under the decreasing-hazard assumption, stronger enforcement leaves adult-tier demand less responsive in proportional terms, so the marginal circumvention cost per teen displaced rises. Teen welfare thus increases as long as this marginal cost is below the marginal benefit, which is the elimination of the behavioral bias. The bound on network feedback ensures that the resulting changes in network size do not cancel out this selection effect.

Welfare and the regulatory penalty. Lastly, we discuss the connection between our previous results and the penalty imposed by the regulator. For the following characterization, we impose Condition 2 and the conditions of Lemma 1. Figure 2 summarizes the equilibrium paths under the ban and safe-tier regimes, at different levels of the regulatory penalty.

For $\tau < \tau_P$, the ban and safe tier regimes coincide with the unregulated equilibrium. Once τ exceeds τ_P, the platform provides the safe tier but does not yet enforce the adult-tier restriction. Usage of the adult tier therefore falls and teen welfare increases discretely to $W_S(0)$, remaining flat until τ_E.

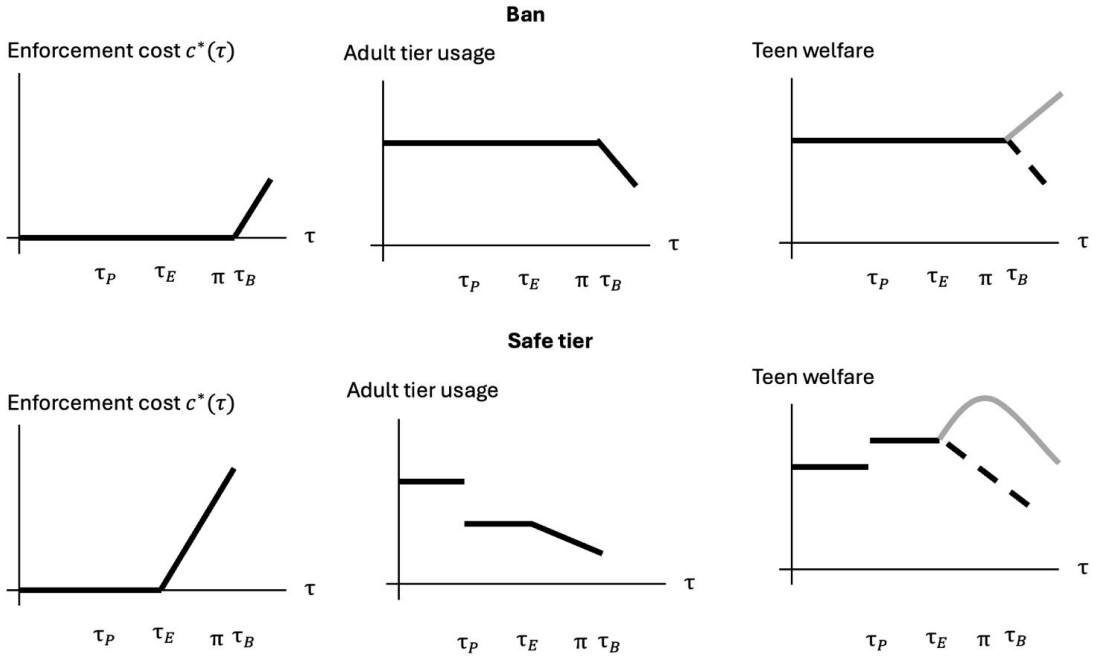
For $\tau \in (\tau_E, \tau_B]$, all outcomes remain flat under the ban, while enforcement increases under the safe-tier regime. Lemma 1 implies that the welfare effect of this additional enforcement depends on the size of the behavioral bias. If the bias is small ($w \leq K_S(0)$), even a small amount of enforcement is socially too costly, so welfare is maximized by providing the safe tier without enforcing the adult-tier restriction. If $w > K_S(0)$, some positive enforcement raises welfare because the benefit from correcting the behavioral bias initially exceeds its social cost. Welfare then increases until this marginal cost exceeds the bias, or continues to increase through τ_B if that point is not reached.

Beyond τ_B, the ban also induces positive enforcement, and the welfare ranking between the two regulatory regimes generally depends on parameters.

To summarize, consider a regulator choosing expected penalties $\tau \leq \tau_B$, the range in which the ban remains unenforced. Over this range, the maximum teen welfare achievable under the safe-tier regime is higher that the welfare achievable under the ban. Under the conditions maintained for this characterization, the welfare-maximizing policy depends on the behavioral

[14] An analogous result holds under the ban. Let $K_B(c) \equiv q_B(c)/|q'_B(c)| + q_B(c)u'(q_B(c))$. If the distribution H has a strictly decreasing hazard rate over the relevant range, then $K_B(c)$ is strictly increasing under Assumption 1. Additionally, Appendix B.2 extends the result to heterogeneous and correlated behavioral wedges.

Figure 2: Effects of the Regulatory Penalty under a Ban and Safe Tier



Notes: The figure summarizes the equilibrium effect of the per-user penalty τ on enforcement, adult-tier use, and teen welfare. The top row corresponds to the ban and the bottom row to the safe-tier regime. Under Condition 2, $\tau_P < (1-\delta)\pi < \tau_E < \pi \leq \tau_B$. Under the ban, enforcement is zero and adult-tier use and teen welfare remain at their unregulated levels through τ_B. Under the safe-tier regime, provision at τ_P causes a discrete fall in adult-tier use and a discrete increase in teen welfare; enforcement begins at τ_E and rises thereafter. Depending on the welfare tradeoff characterized in the text, welfare after τ_E may fall, rise and then fall, or continue to rise through τ_B.

bias relative to marginal social cost. If the bias is small ($w \leq K_S(0)$) the regulator should set the penalty high enough to induce provision of the safe tier, but below the level at which the platform begins to enforce the age restriction. Otherwise if the bias is high enough ($w > K_S(0)$), the optimal penalty exceeds τ_E and induces both safe-tier provision and positive enforcement.

3 Age Restrictions through the Lens of our Model

Under the ban implemented in Australia, social media platforms must take reasonable steps to prevent individuals under 16 from holding accounts. As discussed above, this restriction requires platforms to remove profitable customers and teens to give up access to their relevant online network. Three months after the ban's implementation, account ownership among 10–15-year-olds had fallen from approximately 52% to 42%, but daily use changed little. Many children who kept their accounts reported that platforms had not checked their age.[15] Our model offers one explanation for this limited response: platforms may find that the expected penalties do not compensate for the revenue lost due to the enforcement of the age restriction,

[15]See eSafety Commissioner's three-month evaluation. Note that children can access some platform content without holding an account, so account ownership and use need not move together.

while teens have reasons to remain in the platform when their friends continue using it.

Canada and the EU have proposed allowing younger users to keep access under certain conditions. Canada's proposed Safe Social Media Act would restrict accounts for under-16s on designated services, but allow the regulator to exempt services that provide adequate safeguards for children.[16] The European Commission's September 2026 KIDS Act proposal would prohibit social media accounts for children under 13, allow supervised accounts with restricted features and a one-hour daily limit for those aged 13–14, and permit independently managed accounts from age 15, subject to protections for minors.[17] Our model suggests that both proposals give platforms incentives to offer a safer product: doing so allows them to continue serving users who would otherwise be excluded. Naturally, whether teens would choose that product also depends on the fraction of their relevant network that they can keep interacting with. For example, the EU proposal requires parental approval of contacts. If these requirements prevent teens from communicating with their relevant network, switching to the restricted account would remain socially costly, as in our extension with imperfect interoperability.

California's Protecting Our Kids from Social Media Addiction Act restricts personalized "addictive feeds" for minors without parental consent, but excludes direct private communications from the definition of an addictive feed.[18] Teens can therefore use communication features without receiving the restricted feed, similar to the safe tier in our model. The experience with alternative feeds in Europe shows that implementation also matters. As discussed above, Meta allowed users to select a non-profiled feed, but this choice could be undone when they navigated elsewhere or reopened the app. More recently, a Dutch court ordered Meta to make the choice persistent and easier to access. The European Commission also preliminarily found that TikTok's time-management tools did little to address the risks of excessive use because users could easily dismiss them.[19] These cases help explain why the regulator must specify and audit the restrictions on the safe tier. Even when an alternative feed works as intended, making it available leaves teens free to choose the ordinary feed, including those who prefer it because of the behavioral bias.

The August 2026 settlement between Meta and U.S. states goes further in prescribing how teenagers may use Facebook and Instagram. It requires default daily time limits and overnight restrictions, but allows teens to continue messaging. It also requires age-assurance measures and an independent auditor to review compliance.[20] Preserving messaging allows teens to remain in contact with their friends after reaching a consumption limit. It also allows Meta to keep these users on the platform, which is the source of the enforcement incentive in our

[16] See Bill C-34, particularly Sections 27–29 of the proposed Digital Safety Act.

[17] See the European Commission's September 17, 2026 announcement and description of the proposal.

[18] See SB 976, Sections 27000.5 and 27001.

[19] See the Amsterdam Court of Appeal's March 2026 decision and the Commission's February 2026 preliminary findings.

[20] See the settlement agreement, Sections II and III.

model when retained users continue generating revenue. The agreement makes stronger time restrictions *conditional* on comparable measures being adopted by competitors, including TikTok and YouTube. This condition is consistent with our competition extension: a platform has less incentive to enforce when displaced users can move to an unrestricted rival instead of remaining on the platform's safe tier.

4 Conclusion

Age restrictions ask platforms to remove profitable customers and teens to give up access to their online network. We study an alternative that allows teens to use a safe tier with restricted features while remaining connected to their friends. The platform can then retain teens displaced from the adult tier as lower-margin customers, reducing its private cost of enforcement. Our model shows that this change can induce provision and enforcement at penalties that leave a ban unenforced, and improve teen welfare.

The practical question is which features a safe tier should restrict and how regulators can verify that these restrictions work while preserving access to the relevant network. A regulator could also mandate safe-tier provision, although the platform's incentives to enforce the adult-tier restriction would still matter. The policies discussed above provide opportunities to study these choices empirically, including whether teens remain on the restricted product or move to less regulated alternatives.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius, "A Model of Online Misinformation," *The Review of Economic Studies*, 2024, *91* (6), 3117–3150.

Allcott, Hunt, Benjamin B. Lockwood, and Dmitry Taubinsky, "Regressive Sin Taxes, with an Application to the Optimal Soda Tax," *The Quarterly Journal of Economics*, 2019, *134* (3), 1557–1626.

_ , Matthew Gentzkow, and Lena Song, "Digital Addiction," *American Economic Review*, 2022, *112* (7), 2424–2463.

_ , Sendhil Mullainathan, and Dmitry Taubinsky, "Energy Policy with Externalities and Internalities," *Journal of Public Economics*, 2014, *112*, 72–88.

Bobrowsky, Meghan, "Meta's Day of Reckoning," *The Wall Street Journal*, August 2026. Accessed 2026-09-03.

Bursztyn, Leonardo, Angela L. Duckworth, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, Christopher Roth, and Cass R. Sunstein, "Tipping Points and

Australia's Social Media Ban," Working Paper 35162, National Bureau of Economic Research April 2026. Revised July 2026.

_ , **Benjamin R. Handel, Rafael Jiménez-Durán, and Christopher Roth**, "When Product Markets Become Collective Traps: The Case of Social Media," *American Economic Review*, 2025, *115* (12), 4105–4136.

_ , **Matthew Gentzkow, Rafael Jiménez-Durán, Aaron Leonard, Filip Milojević, and Christopher Roth**, "Measuring Markets for Network Goods," *Review of Economic Studies*, 2026.

Crémer, Jacques, Patrick Rey, and Jean Tirole, "Connectivity in the Commercial Internet," *Journal of Industrial Economics*, 2000, *48* (4), 433–472.

Deneckere, Raymond J. and R. Preston McAfee, "Damaged Goods," *Journal of Economics & Management Strategy*, 1996, *5* (2), 149–174.

eSafety Commissioner, "Early Days, Early Insights: Understanding Experiences of Social Media Age Restrictions at the Three-Month Follow-Up," Australian Government, Office of the eSafety Commissioner July 2026. First report of a two-year longitudinal evaluation; $n = 803$ children aged 10–15 and their parents, baseline November–December 2025, follow-up March–April 2026. Accessed 2026-09-04.

_ , "Social Media Minimum Age: Compliance Update," Australian Government, Office of the eSafety Commissioner March 2026. First compliance report on the Social Media Minimum Age obligation. Reports approximately 4.7 million accounts removed, deactivated or restricted by mid-December 2025 and more than 310,000 further accounts by March 2026. Accessed 2026-09-04.

_ , "Social Media Minimum Age: Three-Month Evaluation Report," 2026. Australian Government, Office of the eSafety Commissioner. First report of a two-year longitudinal evaluation; n = 803 children aged 10–15 and their parents, baseline December 2025, follow-up March–April 2026.

Farrell, Joseph and Garth Saloner, "Standardization, Compatibility, and Innovation," *RAND Journal of Economics*, 1985, *16* (1), 70–83.

Hua, Xinyu and Kathryn E. Spier, "Holding Platforms Liable," *American Economic Journal: Microeconomics*, 2025, *17* (4), 68–101.

Katz, Michael L. and Carl Shapiro, "Network Externalities, Competition, and Compatibility," *American Economic Review*, 1985, *75* (3), 424–440.

Köhler-Dauner, Franziska, Lena Peter, Emily Sitarski, Katrin Chauviré-Geib, Ann-Christin Haag, and Jörg M. Fegert, "Digital Child Protection in Social Networks: Age Verification and Age-Tiered Regulation in Europe," *Child and Adolescent Psychiatry and Mental Health*, 2025, *19* (1), 143.

Kraakman, Reinier H., "Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy," *The Journal of Law, Economics, and Organization*, 1986, *2* (1), 53–104.

Lefouili, Yassine and Leonardo Madio, "The Economics of Platform Liability," *European Journal of Law and Economics*, 2022, *53* (3), 319–351.

Mussa, Michael and Sherwin Rosen, "Monopoly and Product Quality," *Journal of Economic Theory*, 1978, *18* (2), 301–317.

O'Donoghue, Ted and Matthew Rabin, "Optimal Sin Taxes," *Journal of Public Economics*, 2006, *90* (10–11), 1825–1849.

Rechtbank Amsterdam, "Bits of Freedom v. Meta Platforms Ireland Ltd.," Judgment in preliminary relief proceedings, District Court of Amsterdam October 2025. Judgment of 2 October 2025, upheld on appeal 28 October 2025; finding breaches of Articles 25, 27(3) and 38 of the Digital Services Act.

Rosaia, Nicola, "Competing Platforms and Transport Equilibrium," *Econometrica*, 2025, *93* (6), 2235–2271.

A Proofs

A.1 Equilibrium and Preliminary Results

First we show that the participation equilibria are unique and highlight several comparative statics. Under the unregulated adult-only benchmark, define the participation map:

$$\Phi_U(q) = 1 - H(-u(q)).$$

Since

$$h(x) = \int f(x - w - \theta)\, dG(\theta),$$

the density h of H satisfies $\sup_x h(x) \leq \sup_x f(x)$ and $h(x) > 0$ for every finite x. Hence Assumption 1 implies $\sup_q |\Phi_U'(q)| < 1$.

Under an adult-only product offering with enforcement c, the participation map is

$$\Phi_B(q; c) = 1 - H(c - u(q)),$$

which satisfies the same derivative bound. When the safe tier is offered, a teen participates if and only if $u(q) + \varepsilon + \max\{0, w + \theta - c\} \geq 0$. The corresponding participation map is

$$\Phi_S(q; c) = 1 - \int F\left(-u(q) - \max\{0, w + \theta - c\}\right) dG(\theta).$$

Its derivative satisfies:

$$\Phi_S'(q; c) = u'(q) \int f\left(-u(q) - \max\{0, w + \theta - c\}\right) dG(\theta) \leq \max_{q \in [0,1]} u'(q) \sup_x f(x) < 1.$$

Thus all three participation maps are contractions and their fixed points are unique.

At $c = 0$, the safe-tier participation map is strictly above the adult-only map: for every q, a positive mass of teens with $\theta < -w$ joins only when the safe tier is available. Thus $\Phi_S(q_0; 0) > q_0$. Since $\Phi_S(q; 0) - q$ is strictly decreasing, its fixed point satisfies $\tilde{q}_0 > q_0$. Moreover, $\tilde{q}_{S0} = G(-w)\left[1 - F(-u(\tilde{q}_0))\right] > 0$.

Because C is strictly increasing and convex on $[0, \infty)$, $C(c) \to \infty$ as $c \to \infty$. Since demand is bounded, platform profit tends to $-\infty$ as $c \to \infty$ for every fixed τ, so there exists an optimal enforcement. Under the strict concavity assumed in the main text, the optimum is unique. Moreover, for $\tau \in [0, T]$, profit at $c = 0$ is at least $-T$, whereas profit at any c is at most $\pi - C(c)$. Every optimum therefore satisfies $C(c) \leq \pi + T$. Optimal enforcement thus lies in a common compact interval and, whenever unique, varies continuously with τ.

We next give the primitive sufficient condition for Condition 1 referenced in the main text. Let $\kappa \equiv \sup_x f(x) \max_{q \in [0,1]} u'(q) < 1$. At any fixed network size q, the increase in participation

generated by making the safe tier available at $c = 0$ is

$$\begin{aligned}\Phi_S(q;0) - \Phi_U(q) &= \int_{-\infty}^{-w} \left[F(-u(q) - w - \theta) - F(-u(q))\right] dG(\theta) \\ &\leq \sup_x f(x) \int_{-\infty}^{-w} (-w - \theta)\, dG(\theta).\end{aligned}$$

Using the contraction bound,

$$\tilde{q}_0 - q_0 \leq \frac{\sup_x f(x)}{1-\kappa} \int_{-\infty}^{-w} (-w - \theta)\, dG(\theta). \tag{9}$$

Moreover,

$$\tilde{q}_{S0} = G(-w)\left[1 - F(-u(\tilde{q}_0))\right] \geq G(-w)\left[1 - F(-u(0))\right].$$

Therefore,

$$(1-\delta)G(-w)\left[1 - F(-u(0))\right] > \frac{\sup_x f(x)}{1-\kappa} \int_{-\infty}^{-w} (-w - \theta)\, dG(\theta) \tag{10}$$

implies Condition 1.

A.2 Ban

We first derive how enforcement affects participation under the ban; this demand response determines the platform's incentive to enforce and the resulting welfare effects. Differentiating the adult-only fixed point gives:

$$q_B'(c) = -h\big(c - u(q_B(c))\big)\left[1 - u'(q_B(c))q_B'(c)\right].$$

Hence

$$q_B'(c) = -\frac{h\big(c - u(q_B(c))\big)}{1 - h\big(c - u(q_B(c))\big)u'(q_B(c))} < 0, \tag{11}$$

where the denominator is positive by Assumption 1.

Proof of Proposition 1. For $\tau \leq \pi$, the coefficient $\pi - \tau$ in Equation (1) is nonnegative. Since $q_B(c)$ is decreasing in c and $C(c)$ is strictly increasing, profit is strictly decreasing in c. Thus $c_B^*(\tau) = 0$ throughout this range.

For $\tau \geq \pi$, the ban enforcement problem is strictly concave by the regularity condition imposed in the main text. Its right derivative at the boundary $c = 0$ is

$$\Pi_{B,c}(0;\tau) = (\pi - \tau)q_B'(0) - C'(0) = (\tau - \pi)|q_B'(0)| - C'(0).$$

Strict concavity implies that the unique optimum is $c = 0$ if this derivative is weakly negative and is strictly positive if it is positive. The derivative crosses zero at $\tau_B = \pi + \frac{C'(0)}{|q_B'(0)|}$. Hence

$c_B^*(\tau) = 0$ for $\tau \leq \tau_B$ and $c_B^*(\tau) > 0$ for $\tau > \tau_B$. Since $C'(0) \geq 0$, $\tau_B \geq \pi$, with equality when $C'(0) = 0$.

Let $\tau_2 > \tau_1 > \tau_B$. At the optimum under τ_1,

$$\Pi_{B,c}\big(c_B^*(\tau_1); \tau_2\big) = (\tau_2 - \tau_1)\left|q_B'\big(c_B^*(\tau_1)\big)\right| > 0.$$

Strict concavity therefore implies $c_B^*(\tau_2) > c_B^*(\tau_1)$. Continuity of profits, uniqueness of the optimum, and $C(c) \to \infty$ imply continuity of optimal enforcement, including at τ_B. ■

For completeness, we derive the welfare effect of direct enforcement under the ban. At a fixed network size, aggregate perceived surplus is At a fixed network size, aggregate perceived surplus is

$$V_B(q, c) \equiv \mathbb{E}\left[\max\{u(q) + \varepsilon + w + \theta - c, 0\}\right].$$

At the equilibrium network, write $\mathcal{V}_B(c) = V_B(q_B(c), c)$. Since perceived utility exceeds true utility by w for every adult-tier user, $W_B(c) = \mathcal{V}_B(c) - wq_B(c)$. The envelope theorem, evaluated at $q = q_B(c)$, gives $\frac{\partial V_B}{\partial c} = -q_B(c)$ and $\frac{\partial V_B}{\partial q} = q_B(c)u'(q_B(c))$. Applying the chain rule,

$$\begin{aligned}\frac{dW_B(c)}{dc} &= -q_B(c) + q_B(c)u'(q_B(c))q_B'(c) - wq_B'(c) \\ &= |q_B'(c)|\left[w - q_B(c)u'(q_B(c))\right] - q_B(c). \end{aligned} \tag{12}$$

The expression can be positive or negative.

A.3 Safe-Tier Demand and Enforcement

We record two demand comparative statics used below. Define

$$\begin{aligned}\alpha(c) &\equiv \int_{c-w}^{\infty} f\big(c - u(q(c)) - w - \theta\big)\, dG(\theta), \\ \beta(c) &\equiv G(c - w) f(-u(q(c))).\end{aligned}$$

Differentiating Equation (3), the terms capturing switches between tiers cancel because these switches do not change total participation, giving

$$q'(c) = -\alpha(c) + u'(q(c))[\alpha(c) + \beta(c)]q'(c).$$

Hence

$$q'(c) = -\frac{\alpha(c)}{1 - u'(q(c))[\alpha(c) + \beta(c)]} \leq 0. \tag{13}$$

Because

$$\alpha(c) + \beta(c) \leq \sup_x f(x),$$

the denominator is positive by Assumption 1.

Adult-tier use is

$$q_A(c) = \int_{c-w}^{\infty} \left[1 - F\big(c - u(q(c)) - w - \theta\big)\right] dG(\theta).$$

Differentiating yields

$$\begin{aligned} q_A'(c) = & -g(c-w)\left[1 - F(-u(q(c)))\right] \\ & - \left[1 - u'(q(c))q'(c)\right]\alpha(c) \leq 0. \end{aligned} \tag{14}$$

Both derivatives are strictly negative whenever $1 - G(c - w) > 0$. If G has a finite upper endpoint $\bar{\theta}$, both derivatives are zero for $c \geq w + \bar{\theta}$. No such enforcement level is optimal: all participating teens are already on the safe tier, so reducing enforcement slightly lowers costs without a first-order revenue loss. Thus both derivatives are strictly negative at every optimal enforcement choice. By contrast, $q_S'(c) = q'(c) - q_A'(c)$ can be negative because the smaller network can induce existing safe-tier users to leave.

At zero enforcement, $q(0) = \tilde{q}_0$, $q_S(0) = \tilde{q}_{S0}$, and $q_A(0) = \tilde{q}_{A0}$. For the fixed-network comparison in Figure 1, compare the two offerings at the same network size q and enforcement level c. A teen uses the adult tier when the safe tier is available only if $u(q) + \varepsilon + w + \theta - c \geq 0$ and $w + \theta - c \geq 0$. The first condition alone determines adult-tier participation under an adult-only offering. Hence, at fixed (q, c), the set of adult-tier users with the safe tier is a subset of the adult-tier users under the ban.

Proof of Proposition 2. We first derive the provision threshold at zero enforcement. At $c = 0$, Equation (5) gives

$$\Pi_S(0;\tau) - \Pi_B(0;\tau) = \pi\left[\tilde{q}_0 - q_0 - (1-\delta)\tilde{q}_{S0}\right] + \tau\left[q_0 - \tilde{q}_{A0}\right].$$

Let $x \equiv \tilde{q}_0 - q_0$ and $s \equiv \tilde{q}_{S0}$. Since $\tilde{q}_{A0} = \tilde{q}_0 - \tilde{q}_{S0}$,

$$q_0 - \tilde{q}_{A0} = s - x. \tag{15}$$

Condition 1 implies $x < (1-\delta)s$, and therefore

$$q_0 - \tilde{q}_{A0} = s - x > \delta s > 0. \tag{16}$$

The profit difference is strictly increasing in τ and crosses zero once, at $\tau_P = \pi\frac{(1-\delta)s - x}{s - x}$.

Its numerator is positive, so $\tau_P > 0$. Moreover, the fact that $\tilde{q}_0 > q_0$ implies $x > 0$, and

$$(1-\delta)(s-x) - \left[(1-\delta)s - x\right] = \delta x > 0.$$

Hence $\tau_P < (1-\delta)\pi$.

Consider next enforcement conditional on offering the safe tier. Conditional on offering the safe tier,

$$\Pi_S(c;\tau) = \delta\pi q(c) + [(1-\delta)\pi - \tau]\, q_A(c) - C(c).$$

For $\tau \leq (1-\delta)\pi$, both demand coefficients are nonnegative, so profit is strictly decreasing in c. For $\tau \geq (1-\delta)\pi$, the enforcement problem is strictly concave by the condition imposed in Proposition 2. Its right derivative at $c = 0$ is

$$m_S(\tau) \equiv \Pi_{S,c}(0;\tau) = \delta\pi q'(0) + [(1-\delta)\pi - \tau]\, q'_A(0) - C'(0). \tag{17}$$

Because $q'_A(0) < 0$, $m_S(\tau)$ is strictly increasing in τ. At $\tau = (1-\delta)\pi$,

$$m_S((1-\delta)\pi) = \delta\pi q'(0) - C'(0) < 0.$$

The unique value at which the derivative equals zero is

$$\tau_E = (1-\delta)\pi + \frac{C'(0) - \delta\pi q'(0)}{|q'_A(0)|} = \pi - \frac{\delta\pi q'_S(0) - C'(0)}{|q'_A(0)|},$$

where the second equality uses $q'_S(0) = q'(0) - q'_A(0)$.

Strict concavity implies that $c^*_S(\tau) = 0$ for $\tau \leq \tau_E$ and $c^*_S(\tau) > 0$ for $\tau > \tau_E$, conditional on offering the safe tier. Since $q'(0) < 0$ and $C'(0) \geq 0$, $\tau_E > (1-\delta)\pi$.

It remains to connect $\tau_E < \pi$ to Condition 2. At $\tau = \pi$, $\Pi_S(c;\pi) = \delta\pi q_S(c) - C(c)$. Hence Condition 2 is equivalent to the existence of some $c > 0$ satisfying $\Pi_S(c;\pi) > \Pi_S(0;\pi)$. By strict concavity, this holds if and only if the right derivative at zero is positive:

$$\Pi_{S,c}(0;\pi) = \delta\pi q'_S(0) - C'(0) > 0.$$

Using the expression for τ_E above, this is in turn equivalent to $\tau_E < \pi$.

Let $\tau_2 > \tau_1 > \tau_E$. At the conditional optimum under τ_1,

$$\Pi_{S,c}\big(c^*_S(\tau_1);\tau_2\big) = (\tau_2 - \tau_1)\left|q'_A\big(c^*_S(\tau_1)\big)\right| > 0.$$

Strict concavity therefore implies $c^*_S(\tau_2) > c^*_S(\tau_1)$. Continuity of profits, uniqueness of the optimum, and $C(c) \to \infty$ imply continuity of conditional optimal enforcement, including at τ_E.

We can now combine the provision and enforcement choices through τ_B. For $\tau < \tau_P < (1-\delta)\pi$, both the adult-only and safe-tier enforcement problems are maximized at $c = 0$, and the zero-enforcement profit comparison above implies that the adult-only product offering yields higher profit. The platform therefore offers only the adult tier.

At $\tau = \tau_P$, both offerings are still optimized at $c = 0$, and the platform is indifferent between

them. For every $\tau > \tau_P$, the same zero-enforcement comparison gives

$$\Pi_S(0;\tau) > \Pi_B(0;\tau).$$

Proposition 1 establishes that $c_B^*(\tau) = 0$ for all $\tau \leq \tau_B$. Thus, throughout $(\tau_P, \tau_B]$, merely offering the safe tier at $c = 0$ strictly dominates the platform's best adult-only outcome. The global optimum therefore includes the safe tier in this interval. Combining this observation with the conditional enforcement threshold gives zero enforcement for $\tau \in (\tau_P, \min\{\tau_E, \tau_B\}]$ and positive enforcement whenever $\tau_E < \tau \leq \tau_B$.

Under Condition 2, the previous steps yield

$$0 < \tau_P < (1-\delta)\pi < \tau_E < \pi \leq \tau_B.$$

For $\tau > \tau_B$, adult-only enforcement may also be positive, and the relative profitability of the two product offerings is not generally ranked by the maintained assumptions. ■

Under Condition 2, the platform enforces with the safe tier but not under the ban for $\tau \in (\tau_E, \tau_B]$. We can express the width of this interval in terms of demand responses and enforcement costs. From Equation (7), $\pi - \tau_E = \frac{\delta\pi q_S'(0) - C'(0)}{|q_A'(0)|}$. Since $\tau_B - \pi = \frac{C'(0)}{|q_B'(0)|}$, the interval has width

$$\tau_B - \tau_E = \frac{\delta\pi q_S'(0) - C'(0)}{|q_A'(0)|} + \frac{C'(0)}{|q_B'(0)|}.$$

If $C'(0) = 0$, then $\pi - \tau_E = \delta\pi \frac{q_S'(0)}{|q_A'(0)|}$. Using $q'(0) = q_S'(0) + q_A'(0)$ and $q'(0) < 0$, $|q_A'(0)| = q_S'(0) + |q'(0)|$. Thus $q_S'(0)/|q_A'(0)|$ measures the net increase in safe-tier use relative to the fall in adult-tier use. It need not equal the fraction of displaced adult-tier users who switch to S, because the smaller network can also induce existing safe-tier users to exit. Specifically,

$$q_S'(0) = g(-w)\left[1 - F(-u(\tilde{q}_0))\right] + G(-w)f(-u(\tilde{q}_0))u'(\tilde{q}_0)q'(0).$$

The first term captures switches from tier A to S, while the second captures the decline in safe-tier use as the network shrinks. Holding the network fixed makes the second term zero, so the net increase in safe-tier use equals the flow of users switching from A to S.

Proof of Corollary 1. Conditional on participation, a teen chooses the safe tier if and only if

$$u(q) + \varepsilon > u(q) + \varepsilon + w + \theta - c,$$

or equivalently, $\theta < c - w$. Substituting the equilibrium enforcement choice gives $\theta < c_S^*(\tau) - w$. As the level of equilibrium enforcement rises, the cutoff rises. Thus, among teens who continue participating, those with lower θ switch from the adult tier to the safe tier first. Note that other teens may exit the platform altogether rather than switch to the safe tier, so the targeting result

does not imply that the total safe-tier participation increases with enforcement. If $c_S^*(\tau) = w$, the cutoff is zero and a participating teen chooses the safe tier if and only if $\theta < 0$. ■

A.4 Teen Welfare

Let $\mathcal{V}_U$ denote aggregate perceived surplus under the unregulated adult-only equilibrium and let $\mathcal{V}_S(c)$ denote aggregate perceived surplus when the safe tier is offered at enforcement level c. Since perceived and true utility differ by exactly w for every adult-tier user,

$$W_U = \mathcal{V}_U - wq_0, \qquad W_S(c) = \mathcal{V}_S(c) - wq_A(c). \tag{18}$$

Proof of Proposition 3. When $\tau = 0$, the enforcement arguments above imply that both product offerings are optimized at $c = 0$. Condition 1 is therefore necessary and sufficient for the platform to strictly prefer the adult-only offering. We still need to show that provision raises teen welfare.

At $c = 0$, the platform offering the safe tier expands the choice set of teens. In this case, evaluated at the old network q_0, no teen's perceived utility can fall. Further, a positive mass of teens with $\theta < -w$ strictly values the safe option. The resulting increase in participation increases the size of the equilibrium network from q_0 to $\tilde{q}_0 > q_0$, which further raises perceived utility for participants. Thus, $\mathcal{V}_S(0) > \mathcal{V}_U$.

Condition 1 also implies, by Equation (16), $q_0 - \tilde{q}_{A0} > 0$. Using Equation (18), $W_S(0) - W_U = \mathcal{V}_S(0) - \mathcal{V}_U + w\left[q_0 - \tilde{q}_{A0}\right] > 0$ Thus, the platform withholds the safe tier even though offering it at zero enforcement strictly raises teen welfare. ■

Marginal welfare from enforcement. At a fixed network size, aggregate perceived surplus when both tiers are offered is

$$V_S(q, c) \equiv \mathbb{E}\left[\max\{0,\ u(q) + \varepsilon,\ u(q) + \varepsilon + w + \theta - c\}\right].$$

Thus $\mathcal{V}_S(c) = V_S(q(c), c)$ and $W_S(c) = \mathcal{V}_S(c) - wq_A(c)$. Holding the network fixed, an increase in c lowers perceived utility by one unit for every adult-tier user. An increase in q raises perceived utility by $u'(q)$ for every participant. Using the envelope theorem, evaluated at the equilibrium network, we get: $\frac{\partial V_S}{\partial c} = -q_A(c)$ and $\frac{\partial V_S}{\partial q} = q(c)u'(q(c))$. Applying the chain rule,

$$\begin{aligned}\frac{dW_S(c)}{dc} &= -q_A(c) + q(c)u'(q(c))q'(c) - wq_A'(c) \\ &= w|q_A'(c)| - q_A(c) - q(c)u'(q(c))|q'(c)|,\end{aligned}$$

where the second equality uses Equations (13) and (14). This is Equation (8).

Whenever adult-tier demand is positive, $q'_A(c) < 0$, so we can write

$$\frac{dW_S(c)}{dc} = |q'_A(c)| \left[w - K_S(c)\right], \qquad K_S(c) = \frac{q_A(c)}{|q'_A(c)|} + q(c)u'(q(c))\frac{|q'(c)|}{|q'_A(c)|}. \tag{19}$$

This includes every optimal enforcement choice. Under Assumption 2, it also includes every finite $c \geq 0$.

Proof of Lemma 1. Fix network utility at v and define adult-tier demand

$$A(c, v) \equiv \int_{c-w}^{\infty} \left[1 - F(c - v - w - \theta)\right] dG(\theta).$$

Let $\bar{F} \equiv 1 - F$ and $\bar{G} \equiv 1 - G$. Substituting $y = \theta - (c - w)$ and integrating by parts gives

$$A(c, v) = \bar{F}(-v)\bar{G}(c - w) + \int_0^{\infty} f(-v - y)\bar{G}(c - w + y)\, dy. \tag{20}$$

Thus, holding the network fixed, adult-tier demand is a nonnegative mixture of translations of the survival function of θ.

Let $h_G(t) \equiv g(t)/\bar{G}(t)$ denote the hazard rate of θ. Assumption 2 implies that h_G is strictly decreasing on the relevant range. Define a probability measure $\mu_{c,v}$ over $T = c - w + y$ by normalizing the terms in Equation (20). Differentiating under the integral gives

$$\begin{aligned} -\frac{A_c(c, v)}{A(c, v)} &= E_{\mu_{c,v}}[h_G(T)], \\ \frac{A_{cc}(c, v)}{A(c, v)} &= E_{\mu_{c,v}}\left[h_G(T)^2 - h'_G(T)\right]. \end{aligned}$$

Therefore, for the fixed-network marginal cost

$$\mathcal{K}(c, v) \equiv \frac{A(c, v)}{-A_c(c, v)},$$

we have

$$\mathcal{K}_c(c, v) = \frac{\mathrm{Var}_{\mu_{c,v}}[h_G(T)] - E_{\mu_{c,v}}[h'_G(T)]}{E_{\mu_{c,v}}[h_G(T)]^2} > 0. \tag{21}$$

The inequality follows because a decreasing hazard has $h'_G \leq 0$, while strict decrease and the full support of F make the distribution of $h_G(T)$ nondegenerate. Thus, the fixed-network marginal cost $A/(-A_c)$ is strictly increasing in enforcement.

We now account for network adjustment. Fix the remaining primitives and $v_0 \equiv u(0)$, and consider the benchmark $u(q) \equiv v_0$. Then $K_S(c) = \mathcal{K}(c, v_0)$. We first place the enforcement choices for nearby network utility functions in a common compact interval.

Choose $\eta_0 > 0$ with $\eta_0 \|f\|_\infty < 1$. For $\|u'\|_\infty \leq \eta_0$ and $u(0) = v_0$, network utility lies in

$[v_0, v_0 + \eta_0]$. Since h is continuous and positive, $d_0 \equiv \min_{t\in[-v_0-\eta_0,-v_0]} h(t) > 0$. Equation (11) implies $|q_B'(0)| \geq d_0$, so

$$\tau_B \leq \pi + \frac{C'(0)}{d_0} \equiv T_0.$$

Choose $\bar{c}$ such that $C(\bar{c}) > \pi + T_0$. The profit bound established above implies that $c_S^*(\tau_B) < \bar{c}$ for every such u. By Equation (21) and continuity, $m \equiv \min_{c\in[0,\bar{c}]} \mathcal{K}_c(c, v_0) > 0$.

Let $v(c) = u(q(c))$. On this common interval, $|v(c) - v_0| \leq \|u'\|_\infty$, and the contraction bound and the smoothness of f and g give uniform bounds on q' and q'' when $\|u'\|_\infty + \|u''\|_\infty$ is small. In particular, differentiating $q(c) = \Phi_S(q(c); c)$ twice gives

$$q''(c) = \frac{\Phi_{S,cc} + 2\Phi_{S,qc}q'(c) + \Phi_{S,qq}[q'(c)]^2}{1 - \Phi_{S,q}},$$

where the partial derivatives are evaluated at $(q(c), c)$. All numerator terms are bounded, and the denominator is bounded away from zero. Therefore, $v'(c) = u'(q(c))q'(c)$ and $v''(c) = u''(q(c))[q'(c)]^2 + u'(q(c))q''(c)$ converge uniformly to zero as the bound on network feedback tends to zero.

Since $q_A(c) = A(c, v(c))$,

$$\begin{aligned} q_A'(c) &= A_c + A_v\, v'(c), \\ q_A''(c) &= A_{cc} + 2A_{cv}v'(c) + A_{vv}[v'(c)]^2 + A_v\, v''(c), \end{aligned}$$

with the partial derivatives evaluated at $(c, v(c))$. The function A is twice continuously differentiable under the maintained density assumptions. Uniform continuity on compact sets therefore implies that q_A, q_A', and q_A'' converge uniformly to $A(c, v_0)$, $A_c(c, v_0)$, and $A_{cc}(c, v_0)$, respectively. Moreover, $-A_c$ is bounded away from zero on the relevant compact set.

Let $N(c) \equiv q(c)u'(q(c))[-q'(c)]$, so $K_S(c) = \frac{q_A(c)+N(c)}{-q_A'(c)}$. Both N and N' converge uniformly to zero. Differentiating the last ratio then shows that K_S' converges uniformly to $\mathcal{K}_c(\cdot, v_0)$. Thus, there exists $\bar{\kappa}_W > 0$ such that

$$\|u'\|_\infty + \|u''\|_\infty < \bar{\kappa}_W \quad \Longrightarrow \quad \sup_{c\in[0,\bar{c}]} \left|K_S'(c) - \mathcal{K}_c(c, v_0)\right| < m/2.$$

It follows that $K_S'(c) > m/2 > 0$ on $[0, c_S^*(\tau_B)]$. ■

Under the ban, let $t(c) \equiv c - u(q_B(c))$. Equation (11) implies

$$K_B(c) = \frac{q_B(c)}{|q_B'(c)|} + q_B(c)u'(q_B(c)) = \frac{1 - H(t(c))}{h(t(c))}.$$

Moreover,

$$t'(c) = \frac{1}{1 - h(t(c))u'(q_B(c))} > 0$$

by Assumption 1. Hence, if H has a strictly decreasing hazard rate over the relevant range, $K_B(c)$ is strictly increasing.

Finally, Equation (19) gives the welfare characterization stated in the main text. If $w \leq K_S(0)$, welfare weakly decreases from $c = 0$ and strictly decreases for every positive c in the relevant range. If $w > K_S(0)$, welfare initially increases. Since $K_S(c)$ is strictly increasing, there can be at most one $c^W \in (0, c_S^*(\tau_B)]$ satisfying $K_S(c^W) = w$; if it exists, welfare increases before c^W and decreases after it. If no such point is reached, welfare continues to increase through $c_S^*(\tau_B)$.

For the welfare comparison through τ_B, Proposition 1 implies that the ban chooses zero enforcement and therefore delivers W_U for every $\tau \leq \tau_B$. By Proposition 3, any penalty in $(\tau_P, \min\{\tau_E, \tau_B\}]$ induces safe-tier provision and delivers $W_S(0) > W_U$. Under Condition 2, $\tau_E < \tau_B$. Continuity and strict monotonicity of conditional enforcement imply that penalties in $[\tau_E, \tau_B]$ implement every enforcement level in $[0, c_S^*(\tau_B)]$. Thus, if $w > K_S(0)$, some attainable positive enforcement raises welfare above $W_S(0)$, and any welfare-maximizing policy in this range uses positive enforcement. If $w \leq K_S(0)$, welfare is instead maximized by safe-tier provision without enforcement, implemented by any penalty in $(\tau_P, \tau_E]$.

B Extensions

In this section we relax three features of the baseline model. First, we allow the safe tier to preserve the teen's network imperfectly. Second, we allow the behavioral bias wedge to vary across teens and to be correlated with their true value of the adult-tier feed. Third, we study a reduced-form retention channel through which competition can weaken the platform's incentive to enforce.

B.1 Imperfect Interoperability

Perfect interoperability makes membership and network value identical across tiers. Suppose instead that safe-tier utility is $U^S(q) = \eta u(q) + \varepsilon$, where $\eta \in [0, 1]$ is the fraction of network value preserved on the safe tier. Perfect interoperability corresponds to $\eta = 1$. Since the safe tier is not behaviorally distorted, this is also its true utility. Adult-tier utility remains as in the baseline model.

Conditional on participation, a teen chooses the safe tier if and only if $\theta < c - w - (1 - \eta)u(q)$. A teen who switches to the safe tier now loses $(1 - \eta)u(q)$ in network value. This loss makes adherence more costly and means that the choice between tiers also depends on how many teens use the platform.

Let $t(q,c) \equiv c - w - (1-\eta)u(q)$. Equilibrium participation solves

$$q = G(t(q,c))\left[1 - F(-\eta u(q))\right] + \int_{t(q,c)}^{\infty} \left[1 - F(c - u(q) - w - \theta)\right] dG(\theta). \tag{22}$$

At the tier boundary, teens face the same participation cutoff on either tier. Switching tiers does not change total participation, so the boundary terms cancel when we differentiate with respect to q. Participation responds to network size through $\eta u'(q)$ on the safe tier and $u'(q)$ on the adult tier. Since $\eta \in [0,1]$, the derivative is nonnegative and bounded above by $\|f\|_\infty \|u'\|_\infty < 1$ under Assumption 1. The original contraction assumption therefore ensures a unique equilibrium for every η.

Let $q_\eta(c)$ denote this equilibrium and let $q_{S,\eta}(c)$ and $q_{A,\eta}(c)$ denote the two demand terms in Equation (22). Write $t_\eta(c) \equiv t(q_\eta(c), c)$ and define

$$\alpha_\eta(c) \equiv \int_{t_\eta(c)}^{\infty} f(c - u(q_\eta(c)) - w - \theta)\, dG(\theta),$$
$$\beta_\eta(c) \equiv G(t_\eta(c)) f(-\eta u(q_\eta(c))).$$

Let $D_\eta(c) \equiv 1 - u'(q_\eta(c))[\alpha_\eta(c) + \eta\beta_\eta(c)]$, which is positive by the contraction assumption. Differentiating the fixed point gives $q'_\eta(c) = -\alpha_\eta(c)/D_\eta(c) \leq 0$. The tier cutoff satisfies $t'_\eta(c) = 1 - (1-\eta)u'(q_\eta(c))q'_\eta(c) \geq 1$. Consequently,

$$q'_{A,\eta}(c) = -g(t_\eta(c))\left[1 - F(-\eta u(q_\eta(c)))\right] t'_\eta(c) - \alpha_\eta(c)\left[1 - u'(q_\eta(c))q'_\eta(c)\right] \leq 0.$$

Both total and adult-tier use therefore decrease with enforcement, strictly so whenever adult-tier demand is positive. As in the baseline, safe-tier use need not increase, because the smaller network can induce existing safe-tier users to leave.

We next show that the provision and early-enforcement results continue to hold when η is sufficiently close to one. As η goes to 1, the participation maps and their first derivatives converge uniformly on compact enforcement intervals to their baseline counterparts. The common contraction bound implies the same convergence for equilibrium participation, tier demands, and their first enforcement derivatives. Profits are therefore continuous in η. Teen welfare is also continuous, by pointwise convergence outside the indifference sets and the finite first moments of tastes.

As in the baseline, assume that platform profit is strictly concave in enforcement over the penalty ranges considered in Proposition 2. Define $\tau_{P,\eta}$ and $\tau_{E,\eta}$ using Equations (6) and (7), evaluated at the zero-enforcement demands and demand derivatives under η. When η is close enough to one, the cannibalization condition still holds because the inequality is strict at $\eta = 1$. Since total and adult-tier demand decrease with enforcement, the proof of Proposition 2 applies with these demands. Moreover, continuity implies that $\tau_{P,\eta} \to \tau_P$ and $\tau_{E,\eta} \to \tau_E$ as $\eta \to 1$. The

ban threshold τ_B does not depend on η. Under the baseline early-enforcement condition, it follows that $0 < \tau_{P,\eta} < (1-\delta)\pi < \tau_{E,\eta} < \pi \leq \tau_B$ for η sufficiently close to one. Since provision strictly raises welfare at $\eta = 1$, continuity also gives $W_{S,\eta}(0) > W_U$ for η sufficiently close to one.

Imperfect interoperability raises the enforcement needed for exact targeting. At $c = w$, some teens harmed by the feed may remain on the adult tier because switching would cost them too much network value. Exact targeting instead requires $c_\eta = w + (1-\eta)u(q_\eta(c_\eta))$. This level exists and is unique because the left-hand side increases with enforcement while the right-hand side weakly decreases. If the platform's enforcement reaches c_η, participating teens choose the safe tier if and only if $\theta < 0$.

B.2 Heterogeneous Bias

We next allow the behavioral wedge $w \geq 0$ to vary across teens, so the pair (θ, w) may be correlated. We maintain independence of ε from (θ, w) and assume that the joint distribution of (θ, w) is absolutely continuous with finite first moments. We also impose the baseline smoothness conditions on the density of the perceived feed value for the enforcement-threshold formulas.

Let $z \equiv \theta + w$ denote the perceived additional value of the adult tier, and let J and j denote its CDF and density. We assume $0 < J(0) < 1$, the analogue of the baseline interiority condition. Perceived utilities become $U^S(q) = u(q) + \varepsilon$ and $U^A(q, c) = u(q) + \varepsilon + z - c$. True adult-tier utility is $u(q) + \varepsilon + \theta - c = u(q) + \varepsilon + z - w - c$. Conditional on participation, a teen chooses the safe tier if and only if $z < c$.

The platform-side results therefore do not require a common bias wedge. Under the ban, participation depends only on $\varepsilon + z$. When both tiers are offered, equilibrium participation solves

$$q = \underbrace{J(c)\big[1 - F(-u(q))\big]}_{q_S(c)} + \underbrace{\int_c^\infty \big[1 - F(c - u(q) - z)\big] dJ(z)}_{q_A(c)}. \tag{23}$$

Thus the platform's demand and profit problem depends on the joint distribution of (θ, w) only through the distribution of z. After redefining the unregulated and zero-enforcement safe-tier quantities using Equation (23), the ban, provision, and enforcement arguments still apply, provided the analogue of Assumption 1 holds, the relevant enforcement problems are strictly concave, and the analogue of Condition 1 holds.

What changes is targeting and welfare. Enforcement sorts teens by perceived adult-tier value $z = \theta + w$, rather than by true feed value θ alone. The following assumption ensures that higher perceived feed value reflects higher true feed value rather than a larger behavioral bias.

Let $k(w, z)$ denote the joint density of (w, z) induced by the joint density of (θ, w), and define

$$m(z) \equiv E[w \mid z].$$

Assumption 3 (Monotone composition of perceived feed value)**.** The joint density $k(w, z)$ satisfies the following negative dependence condition: for every $w_1 < w_2$ and $z_1 < z_2$,

$$k(w_1, z_1)k(w_2, z_2) \leq k(w_1, z_2)k(w_2, z_1). \tag{24}$$

The marginal density j is positive and strictly log-convex over $z \geq 0$. In addition, j and $b(z) \equiv \int_0^\infty wk(w, z)\, dw$ are twice continuously differentiable, with the derivatives needed below bounded on the relevant range.

The first condition requires that, as perceived feed value rises, the conditional distribution of behavioral bias shifts downward in the likelihood-ratio order. In particular, $m(z)$ is weakly decreasing. Higher perceived feed value therefore reflects higher true feed value rather than greater behavioral bias. The second condition restricts the shape of the density over $z \geq 0$. Log-convexity implies a decreasing hazard rate for J and, holding network utility fixed, ensures that stronger enforcement shifts the composition of marginally displaced teens toward higher values of z.

The first restriction can capture a setting in which teens who are more harmed by the feed also face larger behavioral biases, for example because of addiction. It requires strong negative dependence between true feed value and bias.[21] The platform-side results do not require Assumption 3. The targeting and provision results below use its negative dependence condition; the density shape and smoothness conditions are also used for the single-crossing welfare result. Proposition 5 gives an alternative targeting result based only on bounded heterogeneity.

The next proposition shows how this condition preserves targeting. With heterogeneous bias, teens who perceive less value from the feed need not derive less true value from it. Under the negative dependence condition, however, the distribution of true feed value among teens who are marginal between tiers shifts toward higher values as enforcement increases.

Proposition 4 (Monotone targeting with heterogeneous bias)**.** *Suppose Assumption 3 holds. Enforcement continues to sort teens according to perceived adult-tier value $z = \theta + w$. As enforcement c increases, the distribution of w conditional on $z = c$ (teens at the tier-switching margin) shifts downward, while the distribution of θ conditional on $z = c$ shifts upward, both in the sense of first-order stochastic dominance.*

Proof. Fix $z_2 > z_1$ with $j(z_1), j(z_2) > 0$, and let $p_r(w) = k(w, z_r)/j(z_r)$ denote the conditional density of w given $z = z_r$. Equation (24) implies

$$p_1(w_1)p_2(w_2) \leq p_2(w_1)p_1(w_2) \qquad \text{for } w_1 < w_2.$$

[21] With finite second moments, decreasing $m(z)$ implies $\mathrm{Cov}(w, z) = \mathrm{Cov}(m(z), z) \leq 0$. Since $z = \theta + w$, this requires $\mathrm{Cov}(w, \theta) \leq -\mathrm{Var}(w)$. Thus, with finite second moments, independence of true feed value and nondegenerate bias rules out this restriction.

Let P_r denote the corresponding CDF. Integrating over $0 \leq w_1 \leq a < w_2$ gives $P_1(a)[1-P_2(a)] \leq P_2(a)[1-P_1(a)]$, so $P_2(a) \geq P_1(a)$ for every a. Thus, conditional on higher perceived feed value, behavioral bias is lower in the sense of first-order stochastic dominance. In particular, $m(z)$ is weakly decreasing.

Since $\theta = z - w$, higher perceived feed value and lower bias imply higher true feed value. For any a,

$$\Pr(\theta \leq a \mid z_2) = \Pr(w \geq z_2 - a \mid z_2) \leq \Pr(w \geq z_1 - a \mid z_1) = \Pr(\theta \leq a \mid z_1).$$

Hence the distribution of θ conditional on z_2 first-order stochastically dominates its distribution conditional on z_1.

Finally, the tier-switching margin is $z = c$, and participation at this margin requires $\varepsilon \geq -u(q(c))$. Independence of ε from (w, z) implies that conditioning on participation does not change these conditional distributions. ■

This proposition described the order in which enforcement reaches teens with different true feed values. We now ask whether enforcement can separate teens harmed by the feed from those who benefit, as in the homogeneous-bias benchmark. The next proposition—which does not require Assumption 3— shows that, when differences in bias are bounded, only teens with true feed values close to zero can be assigned to the wrong tier.

Proposition 5 (Approximate targeting with heterogeneous bias)**.** *Suppose there exist $\bar{w} \geq 0$ and $\Delta \geq 0$ such that $|w - \bar{w}| \leq \Delta$ for every teen. At enforcement $c = \bar{w}$, any participating teen with $\theta < -\Delta$ chooses the safe tier, and any participating teen with $\theta > \Delta$ chooses the adult tier. Hence only teens with $|\theta| \leq \Delta$ can be sorted differently from the benchmark rule that assigns the safe tier to teens with $\theta < 0$.*

Proof. If $\theta < -\Delta$, then $\theta + w < -\Delta + \bar{w} + \Delta = \bar{w}$, so the cutoff $z < c$ implies that the teen chooses the safe tier. Conversely, if $\theta > \Delta$, then $\theta + w > \Delta + \bar{w} - \Delta = \bar{w}$, so a participating teen chooses the adult tier. ■

When $\Delta = 0$, all teens have the same bias $\bar{w}$, and the result reduces to the baseline case with homogeneous bias: at $c = \bar{w}$, participating teens choose the safe tier if and only if $\theta < 0$. This argument also applies to discrete distributions of tastes and biases. To induce this allocation, the regulator must be able to choose a penalty τ such that $c_S^*(\tau) = \bar{w}$.

We next derive the welfare effect of enforcement and show how Assumption 3 gives a single-crossing result analogous to the baseline. Define the total behavioral wedge among adult-tier users as $B_A(c) \equiv \int_{A(c)} w \, di$. We can write aggregate true welfare as $W_S(c) = \mathcal{V}_S(c) - B_A(c)$, where $\mathcal{V}_S(c)$ is aggregate perceived surplus. Applying the same envelope argument as in Equation (8) gives $\frac{dW_S(c)}{dc} = -q_A(c) + q(c)u'(q(c))q'(c) - B'_A(c)$. Because the equilibrium network is weakly

decreasing in c, both the cutoff $z \geq c$ and the adult-tier participation condition become harder to satisfy as enforcement rises. Thus the adult-tier user sets are nested. Since $w \geq 0$, this implies $B'_A(c) \leq 0$ wherever the derivative exists. At an enforcement level for which $q'_A(c) < 0$, define the average behavioral wedge among teens marginally displaced from the adult tier by

$$\bar{w}_M(c) \equiv \frac{|B'_A(c)|}{|q'_A(c)|}. \tag{25}$$

Then

$$\frac{dW_S(c)}{dc} = |q'_A(c)|\bar{w}_M(c) - q_A(c) - q(c)u'(q(c))|q'(c)| = |q'_A(c)|\left[\bar{w}_M(c) - K_S(c)\right]. \tag{26}$$

The homogeneous model is the special case $\bar{w}_M(c) = w$. With heterogeneous bias, both the marginal benefit and the marginal social cost of enforcement can in principle vary with c. The next lemma shows that Assumption 3 restores the single-crossing property.

Lemma 2 (Single crossing with heterogeneous bias)**.** *Suppose Assumption 3 holds, the analogue of Assumption 1 holds, and the safe-tier enforcement problem is strictly concave over the penalty range in Proposition 2. Fix $u(0)$ and the remaining primitives, and suppose u is twice continuously differentiable. There exists $\bar{\kappa}_H > 0$ such that, if $\|u'\|_\infty + \|u''\|_\infty < \bar{\kappa}_H$, then $\bar{w}_M(c) - K_S(c)$ is strictly decreasing for $c \in [0, c^*_S(\tau_B)]$.*

Proof. We first hold network utility fixed at some v. Adult-tier demand is

$$A(c, v) = \int_c^\infty \left[1 - F(c - v - z)\right] dJ(z). \tag{27}$$

Let $s(v) \equiv 1 - F(-v)$. Differentiating with respect to enforcement gives

$$-A_c(c, v) = s(v)j(c) + \int_0^\infty f(-v - y)j(c + y)\, dy. \tag{28}$$

Likewise, total behavioral bias among fixed-network adult-tier users is

$$B(c, v) = \int_c^\infty m(z)\left[1 - F(c - v - z)\right] j(z)\, dz,$$

so

$$-B_c(c, v) = s(v)j(c)m(c) + \int_0^\infty f(-v - y)j(c + y)m(c + y)\, dy. \tag{29}$$

Hence the average wedge among teens displaced by a marginal increase in fixed-network enforcement is

$$M(c, v) \equiv \frac{-B_c(c, v)}{-A_c(c, v)}. \tag{30}$$

Divide the numerator and denominator in Equation (30) by $j(c)$ and define $r_c(y) \equiv \frac{j(c+y)}{j(c)}$.

Since j is log-convex, for $c_2 > c_1$ the ratio $\frac{r_{c_2}(y)}{r_{c_1}(y)}$ is weakly increasing in y and equals one at $y = 0$. Thus, if $\nu_{c,v}$ denotes the probability distribution on $y \geq 0$ obtained by normalizing an atom of mass $s(v)$ at $y = 0$ together with density $f(-v - y)r_c(y)$ for $y > 0$, then, for $c_2 > c_1$, $\nu_{c_2,v}$ dominates $\nu_{c_1,v}$ in the monotone likelihood-ratio order.

Equation (30) can be written as $M(c, v) = E_{\nu_{c,v}}[m(c + Y)]$. By Proposition 4, m is weakly decreasing. Therefore, for $c_2 > c_1$,

$$M(c_2, v) = E_{\nu_{c_2,v}}[m(c_2 + Y)] \leq E_{\nu_{c_2,v}}[m(c_1 + Y)] \leq E_{\nu_{c_1,v}}[m(c_1 + Y)] = M(c_1, v),$$

where the second inequality uses the first-order stochastic dominance implied by the monotone-likelihood-ratio ordering. Thus marginal behavioral bias is weakly decreasing in enforcement when network utility is fixed.

We next consider the marginal social cost. Log-convexity of j implies that J has a decreasing hazard rate on $z \geq 0$. Indeed,

$$\frac{1 - J(c)}{j(c)} = \int_0^\infty \frac{j(c + y)}{j(c)}\, dy,$$

which is strictly increasing in c because j is strictly log-convex. Applying the fixed-network argument in the proof of Lemma 1, with the survival function of z in place of $\bar{G}$, therefore gives $\mathcal{K}_c(c, v) > 0$, where $\mathcal{K}(c, v) \equiv A(c, v)/[-A_c(c, v)]$.

Fix $v_0 = u(0)$ and the remaining primitives. The compactness argument in the proof of Lemma 1 applies with the distribution of $\varepsilon + z$ in place of H. Hence a common interval $[0, \bar{c}]$ contains $[0, c_S^*(\tau_B)]$ for all sufficiently small network effects. Continuity gives $\eta \equiv \min_{c \in [0,\bar{c}]} \mathcal{K}_c(c, v_0) > 0$. Since $M(c, v_0)$ is weakly decreasing and continuously differentiable,

$$\frac{d}{dc}\left[M(c, v_0) - \mathcal{K}(c, v_0)\right] \leq -\eta < 0 \tag{31}$$

throughout this interval.

We now account for endogenous network utility, $v(c) = u(q(c))$. The functions q' and q'' are uniformly bounded on $[0, \bar{c}]$, and $v - v_0$, v', and v'' converge uniformly to zero as the bound on network feedback tends to zero.

Write $q_A(c) = A(c, v(c))$ and $B_A(c) = B(c, v(c))$. The chain rule gives $q_A'(c) = A_c + A_v\, v'$, $q_A''(c) = A_{cc} + 2A_{cv}v' + A_{vv}(v')^2 + A_v\, v''$, $B_A'(c) = B_c + B_v\, v'$, and $B_A''(c) = B_{cc} + 2B_{cv}v' + B_{vv}(v')^2 + B_v\, v''$, where the partial derivatives are evaluated at $(c, v(c))$. The finite first moment of w gives $\int b(z)\, dz < \infty$. Together with the boundedness of f and f' and the smoothness of j and b, this makes A and B twice continuously differentiable on the relevant compact set. Therefore the derivatives in the last display converge uniformly to their fixed-network

counterparts. Since $-A_c$ is bounded away from zero, it follows that

$$\frac{d}{dc}\left[\bar{w}_M(c) - K_S(c)\right] \longrightarrow M_c(c, v_0) - \mathcal{K}_c(c, v_0)$$

uniformly on $[0, \bar{c}]$ as $\|u'\|_\infty + \|u''\|_\infty \to 0$.

By Equation (31), the limiting derivative is at most $-\eta$. Choose $\bar{\kappa}_H > 0$ so that the uniform difference from this derivative is less than $\eta/2$. The derivative of $\bar{w}_M - K_S$ is then below $-\eta/2$ throughout $[0, c_S^*(\tau_B)]$. ■

Lemma 2 gives the same qualitative welfare characterization conditional on offering the safe tier. If $\bar{w}_M(0) \leq K_S(0)$, any positive enforcement in the relevant range lowers welfare. If $\bar{w}_M(0) > K_S(0)$, welfare initially rises. If $\bar{w}_M(c) = K_S(c)$ is reached within $[0, c_S^*(\tau_B)]$, the solution is unique and welfare decreases beyond it; otherwise welfare continues to rise over the interval. To compare this conditional welfare optimum with the unregulated outcome, we also need a welfare gain from provision.

The welfare gain from *provision* also survives under Assumption 3. Let B_U denote the total behavioral wedge among adult-tier users in the unregulated equilibrium. Then

$$W_S(0) - W_U = \mathcal{V}_S(0) - \mathcal{V}_U + B_U - B_A(0). \tag{32}$$

Corollary 2 (Safe-tier provision with heterogeneous bias)**.** *Suppose the analogue of Condition 1 and Assumption 3 hold. Then offering the safe tier at zero enforcement strictly raises teen welfare:* $W_S(0) > W_U$.

Proof. Let q_0 denote the unregulated adult-tier use and let $\widetilde{q}_0$ denote total use when both tiers are offered at zero enforcement. Expanding the choice set implies $\widetilde{q}_0 > q_0$. When $c = 0$, all teens with $z < 0$ who choose the adult tier when it is the only product offered instead join the safe tier. Instead, among teens with $z \geq 0$, the larger network will only add more adult-tier users relative to the unregulated benchmark.

Let L represent the mass of adult-tier users with $z < 0$ removed from the adult tier by offering the safe-tier, and let E denote the mass of new adult-tier users with $z \geq 0$ generated by the larger network. Then: $q_0 - \widetilde{q}_{A0} = L - E$. The analogue of Condition 1 implies $q_0 - \widetilde{q}_{A0} > 0$, so $L > E$.

The first part of Assumption 3 implies that $m(z) = E[w \mid z]$ is weakly decreasing wherever the conditional mean is defined. Independence of ε from (w, z) means that selecting users by their membership benefit does not change this conditional mean. Hence the average behavioral wedge among the removed $z < 0$ users is at least $m(0)$, whereas the average wedge among the added $z \geq 0$ adult-tier users is at most $m(0)$. Therefore, $B_U - B_A(0) \geq m(0)(L - E) \geq 0$.

The perceived-surplus term $\mathcal{V}_S(0) - \mathcal{V}_U$ in Equation (32) is strictly positive because the safe tier expands the choice set and equilibrium network. Thus $W_S(0) > W_U$. ■

Finally, keeping the conditions of Corollary 2 and Lemma 2, the analogue of Condition 2, and the ban's strict-concavity condition, we can extend the regulator's welfare comparison through τ_B. The ban chooses zero enforcement throughout this range, so its welfare remains W_U. The safe-tier regime can obtain strictly higher welfare by inducing provision. When $\bar{w}_M(0) \leq K_S(0)$, its optimal policy in this range uses provision without enforcement; when $\bar{w}_M(0) > K_S(0)$, its optimal policy combines provision with positive enforcement and achieves welfare strictly above $W_S(0)$.

B.3 Competition

In this section, we focus on how competition can weaken the platform's enforcement incentive, as modeled through a reduced-form approach of lower retention of the safe-tier margin. We consider how the incumbent's incentive changes when it captures less of the safe-tier margin generated by enforcement. In this extension, we fix the user-side diversion schedules $q_A(c)$ and $q_S(c)$ so the analysis is not a full fledged multi-platform equilibrium case.

Let $\mu \in [0, 1]$ denote the share of the safe-tier revenue generated by enforcement that the incumbent fails to retain. The baseline from the paper corresponds to $\mu = 0$. We hold the demand schedules and network effects fixed. Up to a constant independent of c, the incumbent's profit when both tiers are offered now becomes

$$\Pi_{S,\mu}(c; \tau) = (\pi - \tau)q_A(c) + \delta\pi q_S(0) + (1 - \mu)\delta\pi \left[q_S(c) - q_S(0)\right] - C(c). \tag{33}$$

The constant term normalizes profit at $c = 0$ to its baseline value. Further, μ lowers the revenue weight on changes in $q_S(c)$. Since q_S does not need to increase globally, the interpretation as lost incremental safe-tier revenue applies to platform enforcement changes that increase safe-tier use.

For $\tau \leq [1 - (1 - \mu)\delta]\pi$, we can write profit as $(1 - \mu)\delta\pi q(c) + ([1 - (1 - \mu)\delta]\pi - \tau)\, q_A(c) - C(c) + \mu\delta\pi q_S(0)$. Both demand coefficients are nonnegative, so profit is strictly decreasing in c. For penalties above this range, assume that the modified enforcement problem is strictly concave in c.

Its marginal profit from enforcement at $c = 0$ is

$$\Pi_{S,\mu,c}(0; \tau) = (\pi - \tau)q_A'(0) + (1 - \mu)\delta\pi q_S'(0) - C'(0).$$

The corresponding conditional enforcement threshold is therefore

$$\tau_E(\mu) = \pi - \frac{(1 - \mu)\delta\pi q_S'(0) - C'(0)}{|q_A'(0)|}. \tag{34}$$

The platform chooses zero enforcement for $\tau \leq \tau_E(\mu)$ and positive enforcement for $\tau > \tau_E(\mu)$,

conditional on offering the safe tier. At $\mu = 0$, this is the baseline threshold τ_E. Enforcement begins below π if and only if

$$(1 - \mu)\delta\pi q_S'(0) > C'(0). \tag{35}$$

Under Condition 2, strict concavity implies $\delta\pi q_S'(0) > C'(0)$ and hence $q_S'(0) > 0$. Equation (34) then implies

$$\frac{d\tau_E(\mu)}{d\mu} = \delta\pi \frac{q_S'(0)}{|q_A'(0)|} > 0.$$

Greater leakage therefore raises the conditional enforcement threshold. Since $\Pi_{S,\mu}(0;\tau) = \Pi_S(0;\tau)$, the provision threshold is unchanged. Under cannibalization, the safe tier is still offered for $\tau \in (\tau_P, \tau_B]$, so the interval in which it induces enforcement while the ban does not has length $\max\{\tau_B - \tau_E(\mu), 0\}$. This length weakly decreases with leakage. If leakage is large enough to violate Equation (35), the safe-tier retention channel is no longer strong enough to induce enforcement below π.

At the extreme of $\mu = 1$, the incumbent receives none of the extra safe-tier revenue created by enforcement, so there is no retention advantage of the safe-tier. However, this does not equate a ban since the safe tier still changes user demand, and $q_A'(0)$ need not equal $q_B'(0)$. In particular,

$$\tau_B - \tau_E(1) = C'(0)\left[\frac{1}{|q_B'(0)|} - \frac{1}{|q_A'(0)|}\right],$$

whose sign is not generally determined. Thus, losing the retention channel eliminates enforcement below π, but need not eliminate enforcement before τ_B.

Competition is one possible interpretation of the retention loss summarized by μ. In particular, teens displaced from the incumbent's adult tier may be more likely to remain on its safe tier when doing so preserves their relevant network and moving to a rival does not. However, when a non-adhering rival provides a close substitute for the relevant network, the incumbent may retain less of the safe-tier monetization generated by enforcement.

Lastly, consider that a safe-tier entrant platform does not supply the same interoperable safe product since it does not inherit the incumbent's network. Thus, without perfect interoperability with the incumbent platform, its users may face a communication loss similar to $(1 - \eta)u(q)$ in Appendix B.1.