LEONARDO MADIO

University of Padova

MARTIN QUINN

Rotterdam School of Management

CONTENT MODERATION AND ADVERTISING IN SOCIAL MEDIA PLATFORMS

June 2024 (First version: March 2023)

Marco Fanno Working Papers – 297

Department of Economics and Management "Marco Fanno"
University of Padova
Via del Santo 33, 35123 Padova

Content Moderation and Advertising in Social Media Platforms

Leonardo Madio* Martin Quinn†

This version: June 2024

We study the incentive of an ad-funded social media platform to curb the presence of unsafe content that entails reputational risk to advertisers. We identify conditions for the platform not to moderate unsafe content and demonstrate how the optimal moderation policy depends on the risk the advertisers face. The platform is likely to under-moderate unsafe content relative to the socially desirable level when both advertisers and users have congruent preferences for unsafe content and to over-moderate unsafe content when advertisers have conflicting preferences for unsafe content. Finally, to mitigate negative externalities generated by unsafe content, we study the implications of a policy that mandates binding content moderation to online platforms and how the introduction of taxes on social media activity and social media platform competition can distort the platform's moderation strategies.

Keywords: Advertising; content moderation; social media platforms; toxic content.

*University of Padova, Department of Economics and Management, Via del Santo, 33, 35123 Padova, Italy. Email: leonardo.madio@unipd.it. Other affiliations: CESifo Research Network

†Rotterdam School of Management, Burgemeester Oudlaan 50, 3062 PA Rotterdam. Email: quinn@rsm.nl.

1 INTRODUCTION

When Elon Musk in 2022 took over Twitter, advertisers, brands, and individual users raised concerns about his intention to relax the platform's content moderation policies. Advertisers and brands were mostly concerned about their "brand safety", which is defined as "the set of measures that aim to protect the brand's image from the negative or harmful influence of inappropriate or questionable content on the publisher's site where the ad impression is served"[1]. Immediately after the take-over of the platform, the world's biggest media buyer, GroupM, classified Twitter as a "high-risk platform" for brands[2] and many luxury brands (e.g., Balenciaga) either paused their ad purchases or quit the platform[3]. Similar concerns had emerged against YouTube in 2018, resulting in the exodus of advertisers (the so-called "Adpocalypse")[4]. Likewise, Facebook was accused of failing to create a safe environment for advertisers. Recent studies have shown for example, that advertisers are willing to bid more if they are aware of the information context in which their ads will appear (Ada et al., 2022), that a good match between ads and content induces a higher click-through-rate (Devaux, 2023) and that the success of a brand campaign may depend on the quality of its context, especially for premium brands (Shehu et al., 2020).

In a two-sided market context (Rochet and Tirole 2003), in which advertisers care about the presence of users but the latter received a negative externality — nuisance cost (Anderson and Coate, 2005) — from the presence of the former, it is not a priori clear whether a stricter content moderation policy ultimately benefits or harms both parties.

This paper studies the incentive of an ad-funded social media platform to design its content moderation policy and advertising strategies and their alignment with the incentives of a social planner. We build on the workhorse model of two-sided markets (Rochet and Tirole 2003), where a social media platform mediates interactions between users who consume online content free of charge and advertisers who pay for ad campaigns. The platform hosts safe and unsafe

[1]See See Smartyads.com for a definition.

[2]See Digiday.com, November 14, 2022, 'The world's biggest media buyer GroupM is telling advertisers that Twitter is a high-risk media buy'.

[3]See Grid.news, November 15, 2022, 'Internal Twitter documents show the scope of advertisers' questions about Elon Musk's policies'.

[4]See The New York Times, March 23 2017, 'YouTube Advertiser Exodus Highlights Perils of Online Ads'.

— but not manifestly unlawful — material, with the latter entailing brand safety issues that render advertising via the social media platform less appealing vis-à-vis their outside option (of advertising via cable TV). Users face a nuisance cost when exposed to ads and, unlike advertisers, users may prefer or have an aversion to unsafe content.[5] In the former case, advertisers' and users' preferences are *congruent*, whereas in the latter case, their preferences are *conflicting*.[6]

Our first result relates to the effect of content moderation on the level of advertiser participation and the platform's incentives to curb the presence of unsafe content. For a given price, a higher content moderation intensity has two effects on advertisers: it makes the social media environment safer (*brand safety effect*), and it repels (respectively, attracts) users who are interested in consuming unsafe (respectively, safe) content (*eyeball effect*). Because the platform generates revenues on the advertiser side only, the social media platform internalizes the effect that a higher content moderation intensity has on advertisers and, indirectly, on users. Despite the presence of unsafe content, we identify sufficient conditions for which the platform finds it optimal not to moderate at all. This happens if users have a strong preference for the presence of unsafe content and if advertisers' marginal reputation loss from being associated with such content is limited.

For our second result, we identify the equilibrium advertising price and content moderation strategy using a model that, for simplicity, relies on a uniform distribution of the outside options of advertisers and users. We study how the platform decision depends on the reputation loss advertisers would face if they were exposed to unsafe content. We show how content moderation plays a fundamental role in users' demand. If the (marginal) moderation cost is

[5]Unsafe or toxic content can have a positive effect on user engagement, which stimulates participation on a platform (Beknazar-Yuzbashev et al., 2022). Moreover, users might like the presence of potentially unsafe content for advertisements but non-toxic for themselves. For example, in 2019, the micro-blogging platform Tumblr lost nearly 30 percent of its traffic and almost 99 percent of its market value after banning porn in late 2018. Such a ban was designed to keep "content that is not brand-safe away from ads". See The Verge, March 14, 2019, 'After the porn ban, Tumblr users have ditched the platform as promised'.

[6]We focus on unsafe content, but a broad class of content exists that is safe but potentially harmful to brand reputation. For example, DoubleVerify (see April 25, 2018, 'A Call for Brand Safety in the Social Media Landscape'., a company working in the media sector to make advertising safe, offers solutions to brands to monitor against eleven types of content, including aviation disasters, violence, hate speech, man-made and natural disasters, pornography, profanity, substance abuse, terrorist events, and weapons and vehicle disasters.

low enough, the platform finds it not too costly to accommodate advertisers' requests for a safer environment; therefore, its optimal content moderation choice increases with reputation loss associated with unsafe unmoderated content. Interestingly, the optimal ad price is U-shaped in the size of reputation loss associated with unmoderated unsafe content to reflect the social media platform's marginal gains from moderation. If the (marginal) moderation cost is sufficiently high, any change in content moderation intensity is costly for the platform. We find that the optimal moderation policy is bell-shaped in the reputation loss associated with unmoderated unsafe content, reflecting the platform's marginal gain from moderation. Because advertisers' utility decreases at a faster rate than any increase in the content moderation intensity, their reputation loss gets larger, and, as a result, the optimal ad price decreases in the magnitude of the reputation loss.

We then show that the platform can moderate "too little" or "too much" relative to the level that maximizes the total welfare, where total welfare excludes the gains that users might derive from consuming unsafe content. We find that the platform over-moderates content when users do not care much about the presence of unsafe content, but advertisers would suffer a significant loss from their presence. In this scenario, the platform is biased towards advertisers and carries out excessive content moderation. On the other hand, the platform under-moderates when users' and advertisers' preferences are either too congruent or too conflicting. This is because, in balancing the participation of users and advertisers in its ecosystem, the platform under-moderates when increasing its level of content moderation would generate more nuisance from advertisers that would offset potential gains from making the platform's environment safer for advertisers (and users, depending on their preferences).

We, therefore, find that there is scope for regulation by the social planner since the platform might have too few incentives to moderate unsafe content. We explore two main regulatory tools. First, we study the effects of a binding mandated content moderation policy, assuming that the social planner imposes stricter content moderation obligations than what the platform would do.[7] For example, the Digital Services Act, which became law in 2022, identifies obligations for large online platforms relative to the presence of illegally produced, uploaded, or sold

[7]Often, these policies are "one-size-fits-all" and do not account for existing differences across platforms.

material while safeguarding freedom of speech. In Germany, the 2017 Network Enforcement Act (NetzDG) obligates platforms to remove unlawful content quickly. We, therefore, study the effects of a mandated content moderation policy that induces the platform to raise its moderation intensity above its privately optimal level. We show that the policy induces the platform to react strategically by raising the ad price. Nevertheless, we find that, in the presence of a uniformly distributed opportunity cost of advertisers, the direct effect of a higher moderation intensity compensates for the loss resulting from a price increase. Although advertisers are better off with a mandated content moderation policy, the effect on user surplus and participation is less straightforward because of two potentially opposite effects. First, users benefit (respectively, suffer) from reducing unsafe content depending on whether they like (respectively dislike) unsafe content. Second, users are exposed to a larger number of ads that generate a nuisance cost. The net effect is overall negative if users prefer or do not dislike "too much" the presence of unsafe content.

Second, we study how a more indirect intervention, through the taxation of social media platforms, can induce online platforms to moderate unsafe content. This is relevant because taxes have often been advocated by policymakers to help platforms better internalize the presence of unsafe content and increase their content moderation intensity.[8] However, in multi-sided markets, interdependence across sides can lead to substantial changes in the business strategies employed by the platform's owner (Belleflamme and Toulemonde, 2018; Bourreau et al., 2018; Kind et al., 2010, 2013; Kind and Koethenbuerger, 2018; Tremblay, 2018). We show that a tax based on the number of users or a tax based on advertising revenues induces the platform to bias its strategy toward the opposite side of the market. Taxing advertising revenues reduces the platform's marginal gains from content moderation; thus, it unintentionally lowers the content moderation intensity. In contrast, taxing platforms based on their user size can have ambiguous results on the platform's incentive to moderate unsafe content.

We further extend our analysis to study how competition affects the decisions of social media platforms. We build on a Hotelling setting with two symmetric social media platforms, single-

[8]The Nobel Prize Laureate Paul Romer put forward a similar proposal. The New York Times, 'A Tax That Could Fix Big Tech ', March 6, 2019.

homing users, and multihoming advertisers. We show that as competition for user attention intensifies, attracting users becomes more salient for the ad-funded business of the platform, and the platform can employ two complementary instruments: controlling the number of ads via the pricing instrument and writing content moderation policies. We show that fiercer competition *tends* to induce social media platforms to adopt lax content moderation.

The paper unfolds as follows. In Section 2, we discuss the related literature. In Section 3, we present the model setup. In Section 4, we present the analysis of the baseline model and compare the privately optimal level of content moderation with the level that a social planner would choose to maximize total welfare. In Section 5, we study the effect of mandating stricter content moderation on social media platforms' strategy, advertisers, and users' surplus. Moreover, we also study how digital platforms impact content moderation policies. In Section 6, we extend the model to platform competition and relax the assumption of reputational losses unrelated to user demand. Section 7 provides concluding remarks and some implications for advertisers, brands, and platforms' owners. We also present implications for policymakers willing to ensure that social media platforms fulfill some social responsibilities when unsafe material circulates within their ecosystem.

2 RELATED LITERATURE

Despite recent regulations for online intermediaries (e.g., EU Digital Services Act) and discussions among the marketer's community (e.g., the Adpocalypse on YouTube), research on platforms' incentives to moderate unsafe content is still limited.[9] The closest paper to ours is that of Liu et al. (2022) and Jiménez Durán (2022).

[9]Exceptions are Chen et al. (2011), Casner (2020), Teh (2022), and Jeon et al. (2021). Chen et al. (2011) study how moderation of user-generated content affects creators' incentives to produce high-quality content. Casner (2020) and Teh (2022), focus on platform governance and screening as an instrument to control competition among sellers. We add to these studies by identifying the social media platform incentives and the indirect effect that a mandated screening policy has on the platform's pricing instrument. Jeon et al. (2021) study the incentive of a marketplace platform to screen out IP-infringing products and the intended and unintended effects of introducing a liability regime for online intermediaries that induces more screening. They identify conditions for higher screening to negatively affect brand owners' innovation incentives and total welfare.

Liu et al. (2022) study content moderation and technology adoption in a social media platform in the presence of user taste heterogeneity. They show that when a platform finds it optimal to moderate content, a revenue model based on advertising produces more extreme content than a revenue model based on subscription. The opposite emerges when the platform does not moderate content. We differ from their study in that we explicitly model the advertisers' side of the market, whose activity level depends on the platform's content moderation policy and pricing strategy. Because users dislike ads, a *see-saw effect* between the two sides of the market can reduce user participation (and surplus) even if the platform moderates unsafe content, and both advertisers and users prefer moderation. Liu et al. also focuses on the revenue model and its impact on content moderation. Instead, we focus on the interplay between ad prices and content moderation and how this richer set of instruments allows the platform to win advertisers.

Jiménez Durán (2022) studies the incentives of social media platforms to ban users and remove toxic content. The platform monetizes users' eyeballs with ads and trade-offs between users' engagement and both safe and unsafe content. As a result, the platform moderates toxic content to the extent to which it raises advertising revenues. This mechanism is akin to ours, although we micro-found advertising revenues by endogenizing user and advertising decisions to the platform. In two field experiments run on Twitter, Jiménez Durán (2022) looks at the effect of moderating hate speech on user surplus and finds no significant effect, which the author rationalizes by users ignoring the potential side effects of hate speech.

Andres and Slivko (2021) and Jiménez Durán et al. (2022) study the effects of enforcing stricter content moderation on online and offline hate crime on content consumption, respectively. Both studies investigate how the introduction of Germany's NetzDG regulation, equivalent to imposing stricter content moderation intensity, influenced offline and online hatred targeting minorities. Andres and Slivko found that the regulation reduced the intensity and volume of hate speech within tweets addressing sensitive topics such as migration and religion. Jiménez Durán et al. also show that the regulation had a statistically significant negative effect on toxic posts by far-right social media users and on crime against refugees in those areas more exposed to the effects of the policy. Beknazar-Yuzbashev et al. (2022) run a field experiment on Facebook,

Twitter, and YouTube, showing that toxic content drives user engagement. The authors show that a platform faces a trade-off between reducing the extent to which toxic content is displayed to users and lowering content consumption, which can be monetized with ads. Andres et al. (2023) explore the implications of YouTube's Adpocalypse, during which major brands departed from the platform due to brand safety concerns. They find that YouTube's adoption of a stricter content moderation policy as a response to the Adpocalypse led content creators to redirect their efforts towards Patreon, favoring a subscription-based business model over an ad-based one.

More broadly, this paper adds to the literature on user-generated content and media outlet provision (Yildirim et al., 2013; Zhang and Sarvary, 2014; Luca, 2015; de Corniere and Sarvary, 2023).[10] We relate to this literature in that we study the harm online content can cause to brands (Yang et al., 2021) and how it impacts platform governance. Our paper also relates to recent studies on information-sharing behavior and algorithmic curation (Abreu and Jeon, 2020; Acemoglu et al., 2021; Berman and Katona, 2020; Kranton and McAdams, 2020; Mueller-Frank et al., 2022). Whereas these papers consider platform strategies and the diffusion of news, we focus instead on the strategies employed by social media platforms to control the quality dimension of the content.

3 THE MODEL

We consider a monopolist social media platform that connects users who consume all available content on the platform and advertisers who run advertising campaigns on behalf of third-party brands. We assume that content creators exogenously develop content on the platform, which can be either safe or unsafe.[11]

The mass of safe content is normalized to 1, whereas the mass of unsafe content is equal to

[10] More broadly, the paper is also related to the literature on user-generated content, although we do not model the creation of content explicitly by users.

[11] Viral content is generated by a handful of popular content creators (e.g., famous YouTubers and influencers on Instagram), and there is a long tail of creators with limited views. On YouTube, content creators monetize views only if they have reached at least 1,000 subscribers and have streamed at least 4,000 hours in the past 12 months. See YouTube, January 16, 2018, 'Additional changes to YouTube partner'.

$\theta(m) \in [0,1]$, where $m \in [0,1]$ is the content moderation policy of the platform, and $\theta(0) = 1$, $\theta(1) = 0$, and $\theta'(0) < 0$. We assume that $\theta(m)$ is continuous and differentiable.

The platform. The platform generates revenues by charging an advertising price p to advertisers that have joined the website, whose mass is denoted by a. The profit of the platform, net of the moderation cost $C(m)$, is defined as

$$\Pi(m,p) := ap - C(m), \tag{1}$$

which is assumed to be concave in both its arguments.

Typically, moderation costs involve hiring human moderators with fixed capacity and employing automatic first-party or third-party classifiers. Since, in our setting, the amount of content is taken as given, we assume that moderation costs are independent of the amount of content circulating on the platform. Moreover, we also assume that moderation costs are independent of the number of ads displayed to users. This is because once a content is flagged as unsafe, it is removed once and for all, and consequently, the associated cost is not contingent on the number of ads but on the difficulty of identifying it as toxic content, determined by its level of toxicity.

Throughout our analysis, we assume the moderation cost is sufficiently convex, with $C(0) = 0$, $C'(0) = 0$, $C'(m) > 0$, and $C''(m) > 0$. In other words, the more stringent the moderation policy of the platform, the larger its moderation cost. This reflects the fact that it becomes more costly for the platform to devote attention to more controversial content (such as conspiracy theories, hate speech, or non-manifestly unlawful content) that requires higher investments or costly technology, that could take the form of text analysis, or ex-post human verification. For example, according to the UK communication regulator, "the return on investment (in the sense of harm reduction) from the continued expansion of moderator capacity may diminish as extra moderators tackle comparatively less harmful content, as a result of effective automatic prioritization" (OFCOM, 2023, p. 7).

Internet users. A mass of Internet users is heterogeneous in their opportunity cost of joining the social media platform, which we denote by ξ, which is independently and identically distributed

over interval $[0, \overline{\xi}]$ with density function $F(\xi)$ and probability distribution function $f(\xi) > 0$. When joining the platform, a user obtains an intrinsic benefit $u > 0$ from the presence of safe content and a benefit or loss ϕ from unsafe content.[12] We distinguish between two cases. In the first one, users benefit from the presence of unsafe content, $\phi = \phi^+ > 0$, and therefore have conflicting preferences to those of advertisers. In the second one, users obtain a disutility from the presence of unsafe content, $\phi = \phi^- < 0$, and therefore have congruent preferences to those of advertisers. We assume that all users are homogeneous in ϕ. Note that our insights would not change qualitatively if we were to consider two groups that only differ in their preferences for unsafe content. The optimal strategy of the platform would then depend on the preferences of the largest consumer group.[13]

Users join the platform free of charge but are exposed to a number of ads, $a \geq 0$. As in Anderson and Coate (2005), we assume that users face a nuisance cost from advertising on the social media platform, and we denote as $\gamma > 0$ the per-unit nuisance cost. The utility of a user that joins the platform is

$$U(a, m) := \underbrace{u \times 1}_{\text{utility from safe content}} + \underbrace{\phi \times \theta(m)}_{\text{dis/utility from unsafe content}} - \underbrace{\gamma \times a}_{\text{nuisance from ads}} \tag{2}$$

The number of users who join the platform is denoted by n.[14]

Advertisers. There is a mass of advertisers who are heterogeneous in their outside option ω (e.g., advertising via cable TV, for example), which is independently and identically distributed over interval $[0, \overline{\omega}]$ with density function $H(\omega)$ and probability distribution function $h(\omega) > 0$. Each

[12] Note that in the absence of safe content, whose value is normalized to u for a given mass of safe content equal to 1, their utility would be negative if users were to dislike unsafe content. As a result, no interior solution would be present.

[13] To provide further clarity, let us examine the situation involving two distinct groups of users. Let us assume that these groups account for a certain proportion, denoted as $\beta \in (0, 1)$ and $1 - \beta$, respectively, out of the entire user base. Moreover, suppose both groups benefit from having access to safe content. However, one of the groups enjoys the inclusion of unsafe content, whereas the other group experiences discomfort due to its presence. Consequently, when making decisions, the platform considers the overall perception of unsafe content among users to decide about its moderation strategy.

[14] As a simplifying assumption, we have omitted the consideration of direct network effects among social media users. Although the presence of users on social media platforms is crucial, our primary focus lies in understanding the impact of cross-group network effects between users and advertisers on the platform's content moderation decisions. Furthermore, we conjecture that incorporating direct network effects into our model would likely enhance or alleviate any effects that we discover.

advertiser runs at most one ad campaign upon joining the platform and pays the advertising price p. We assume that each ad is displayed only once to users on the platform. For a given mass of users n on the platform, advertisers obtain revenues rn where r reflects the advertiser's revenues per impression. We capture the negative effect of unsafe content on advertisers — brand safety issues — by assuming that their presence renders them less appealing to the social media platform than their outside options. Specifically, the marginal loss for unsafe content— a measure of the brand risk associated with the presence of unsafe content on the platform— is denoted by $\lambda > 0$, which we assume to be exogenously given and homogenous across advertisers.[15] Moreover, we assume that advertisers' reputational loss does not depend on the number of users. For instance, this situation could be exemplified by a scandal, where the reputational losses for brands extend beyond the number of users joining the social media network. We relax this assumption in Section 6.2. The utility of an advertiser that runs its campaign on the platform is

$$V(m,p) := \underbrace{r \times n}_{\text{revenues per impression}} - \underbrace{\lambda \times \theta(m)}_{\text{reputation loss}} - \underbrace{p}_{\text{price}}. \tag{3}$$

Timing. The timing of the game is as follows. In the first stage, the social media platform decides its ad price, $p > 0$, and the content moderation policy, $m \in [0,1]$. In the second stage, users and advertisers form fulfilled expectations regarding the number of advertisers and users joining the social media platform.

[15] This assumption requires further clarification. For the sake of simplicity and ease of analysis, we assume that the parameter λ remains constant across advertisers, while their heterogeneity lies in their outside options. Alternatively, if we were to assume heterogeneity in λ among advertisers, the analysis would become computationally challenging. Another approach would involve considering two distinct categories of brands: those that exhibit greater risk aversion towards unsafe content (e.g., clothing brands) and those that are less risk-averse (e.g., online betting brands) when it comes to associating with such content. If the platform's moderation policy does not depend on the advertiser type, in this scenario, the platform would need to adjust the degree of content moderation to account for the varying impacts on different brand types. Overall, introducing heterogeneity among brands in the context of non-discriminatory content moderation would only make the analysis more nuanced.

4 ANALYSIS

This section outlines a potential trade-off between access to a broader audience and brand safety. Then, we solve the model to identify the platform's equilibrium price and content moderation.

4.1 A simple trade-off for advertisers and the platform

As discussed, advertisers recently raised several concerns about the lax content moderation policy that major platforms carry out. Yet, stricter content moderation may not necessarily benefit advertisers. To understand why, let us first determine the level of activity on the platform for a given price. Formally, the masses of users and advertisers are probabilities such that $a(n, m, p) = \Pr(V \geq 0) = H(rn(a, m) - \lambda\theta(m) - p)$ and $n(a, m) = \Pr(U \geq 0) = F(u + \phi\theta(m) - \gamma a(n, m, p))$.

Due to the feedback loop between users and advertisers, solving for a fixed point is useful. We can write the two masses as follows:

$$a(m, p) = H(rF(u + \phi\theta(m) - \gamma a(m, p)) - \lambda\theta(m) - p),$$

$$n(m, p) = F(u + \phi\theta(m) - \gamma H(rn(m, p) - \lambda\theta(m) - p)),$$

Differentiating $a(m, p)$ with respect to m, we have $\frac{da(m,p)}{dm} = \frac{dH(\cdot)}{dm}$. Dropping the arguments for ease of notation yields

$$\begin{aligned}\frac{dH(\cdot)}{dm} =& -\lambda\theta'(m)h(\cdot) + h(\cdot)rf(\cdot)\Big(\phi\theta'(m) - \gamma\frac{dH(\cdot)}{dm}\Big) \\ =& \frac{h(\cdot)\theta'(m)\Big(rf(\cdot)\phi - \lambda\Big)}{1 + \gamma rf(\cdot)h(\cdot)}.\end{aligned}$$

The sign of $\frac{da(m,p)}{dm}$ depends on two main components. First, a (positive) *brand safety effect* because a higher moderation intensity reduces the advertisers' reputation loss of being associated with unsafe content. The sign of this effect is captured by $h(\cdot)\theta'(m)\lambda > 0$. Second, an

eyeball effect is associated with a change in user demand. The sign of this effect is captured by $h(\cdot)rf(\cdot)\phi\theta'(m)$ and depends on whether ϕ is positive or negative. In other words, a stricter moderation policy can lead to more or fewer users' participation on the platform depending on whether they suffer (i.e., $\phi = \phi^- < 0$) or draw positive utility (i.e., $\phi = \phi^+ > 0$) from the presence of unsafe content on the platform. Because $\theta'(m) < 0$, in the former case, the eyeball effect is positive, which is sufficient to ensure that more advertisers will join the platform.[16] In the latter case, the eyeball effect is negative, which means that the audience and the reputation of the advertisers have a trade-off. The net effect is positive if the brand safety effect is more prominent than the eyeball effect, whereas it is negative otherwise.

The following lemma presents conditions for an increase in m to raise the participation level of advertisers to the platform.

Lemma 1. *For any given price, stricter content moderation leads to more advertising if* $\lambda > rf(\cdot)\phi$. *This is always the case if users and advertisers have congruent preferences, i.e.,* $\phi = \phi^- < 0$.

The trade-off between reaching a broader audience and keeping advertisers safe is also in the social media platform. For a given price, differentiating the platform's profit with respect to m yields

$$\frac{\partial \Pi(m,p)}{\partial m} = p\frac{da(m,p)}{dm} - C'(m) = p\frac{dH(\cdot)}{dm} - C'(m).$$

It follows that a sufficient condition for the platform not to have any incentive to engage in content moderation if the first-order condition is negative at $m = 0$. Because $C'(0) = 0$, this case arises if $\frac{da(m,p)}{dm} < 0$ that is, if $\lambda < rf(\cdot)\phi$, which is only possible if users' preferences strongly conflict with those of advertisers. The intuition is that, in this case, because the platform needs users to attract advertisers and monetize eyeballs, it has to sacrifice advertisers' safety.

In all other cases, the platform's content moderation policy, defined at equilibrium as $m^\star$, is in $(0, 1]$. The intuition is quite simple: for a given price, the platform has the incentive to moderate (at least partially) unsafe content as long as this brings additional advertising revenues. The following proposition summarizes this discussion.

[16] A sufficient condition for $\frac{dH(\cdot)}{dm} > 0$ is that $rf(\cdot)\phi - \lambda < 0$, which is always the case if $\phi < 0$.

Proposition 1. *For any given positive ad price, if $\lambda < rf(\cdot)\phi$, the platform does not moderate unsafe content and chooses $m^\star = 0$. In all other cases, the platform's moderation policy is $m^\star \in (0, 1]$.*

The next section characterizes the platform's optimal ad price and moderation policy. For simplicity, we assume that the outside options of the advertisers and users are distributed uniformly.

4.2 Optimal level of content moderation

In this section, we provide a closed-form solution for our analysis. Formally, we assume the following:

$$\xi \sim \mathcal{U}[0,1] \qquad \omega \sim \mathcal{U}[0,1] \qquad (A1)$$

$$C(m) = \frac{cm^2}{2}, \quad c > \frac{(\lambda - \phi r)^2}{2(\gamma r + 1)} \qquad (A2)$$

$$\theta(m) = 1 - m \qquad (A3)$$

$$\gamma + \phi^+ < u < 1 - \phi^+ \qquad \gamma < 1 - 2\phi^+ \qquad (A4)$$

$$\frac{\lambda}{u + \phi} < r < 1 \qquad u + \phi > \lambda \qquad (A5)$$

(A1) assumes that the outside options of users and advertisers are both uniformly distributed in $[0, 1]$. This means that $f(\cdot)$ and $h(\cdot)$ are equal to 1. (A2) states that the moderation cost is quadratic and c, the cost parameter, is sufficiently high to ensure that the platform's profit is concave in m and p. (A3) implies that the amount of unsafe content decreases linearly with the moderation intensity of the platform. This assumption, which greatly simplifies the analysis, can be consistent with the fact that, besides algorithmic checking, several social media platforms employ human moderators that carry out manual checking. (A4) and (A5) ensure that the demand for users and advertisers is always positive and, therefore, the market is never fully covered.

The program of the platform under assumptions (A1-A5) is

$$\max_{m,p} \Pi(m,p) = pa(m,p) - C(m) = p\frac{ru - p - (1-m)(\lambda - r\phi)}{1+\gamma r} - \frac{cm^2}{2}.$$

The following corollary presents the equilibrium ad price and content moderation from the simultaneous decision of the platform.

Corollary 1. *Under Assumptions (A1-A5), the platform sets the following content moderation policy and price:*

(i) If $\lambda \leq \phi r$*:*

$$m^\star = 0 \qquad p^\star = \frac{r(u+\phi) - \lambda}{2}.$$

(ii) If $\phi r < \lambda < \phi r + \frac{2c(\gamma r+1)}{ru}$*:*

$$m^\star = \frac{(\lambda - r\phi)(r(u+\phi) - \lambda)}{2c(\gamma r+1) - (\lambda - r\phi)^2} \in (0,1) \qquad p^\star = \frac{c(\gamma r+1)(r(u+\phi) - \lambda)}{2c(\gamma r+1) - (\lambda - r\phi)^2}. \tag{4}$$

(iii) If $\lambda \geq \phi r + \frac{2c(\gamma r+1)}{ru}$*:*

$$m^\star = 1 \qquad p^\star = \frac{ru}{2}.$$

If consumers have conflicting preferences for content moderation and the marginal reputation loss of advertisers for being associated with unsafe content is low enough, the platform moderates no content. As consumers gain from the presence of unsafe content, ad price increases in ϕ. Also, as more users join the platform, the eyeball effect grows larger than the brand safety effect. An interior moderation level is present for $\phi r < \lambda < \phi r + \frac{2c(\gamma r+1)}{ru}$, a case that exists both when preferences are congruent and when they are conflicting. In this case, the risk carried out by the presence of unsafe content is more significant for advertisers. Therefore, increasing the intensity of content moderation positively affects advertisers' participation and the platform's monetization incentives. Thus, the platform finds it optimal to engage in a partial content moderation $m^\star \in (0,1)$. Finally, the platform finds it optimal to engage in full content moderation if advertisers' losses from their association with unsafe content are large enough. The ad price becomes independent of users' preferences for moderation and only reflects the economic

value ru attached to safe content. In the rest of the analysis, we focus on the most interesting case: partial content moderation of unsafe content, i.e., $m^\star \in (0,1)$.

How does brand risk affect the platform's strategy? To better understand how the optimal price and the moderation strategy of the platform depend on the advertisers' aversion to unsafe content, in what follows, we perform simple comparative statics of $m^\star$ and $p^\star$ with respect to λ, the marginal reputation loss associated with the presence of unsafe content. We restrict our attention to (ii) in Corollary 1, therefore focusing on the interior content moderation solution. Differentiating (4) with respect to λ yields

$$\begin{aligned}\frac{\partial p^\star}{\partial \lambda}\Big|_{m=m^\star\in(0,1)} =& \frac{c(\gamma r+1)((\lambda-r\phi)(r(2u+\phi)-\lambda)-2c(\gamma r+1))}{(2c(\gamma r+1)-(\lambda-r\phi)^2)^2},\\ \frac{\partial m^\star}{\partial \lambda}\Big|_{m=m^\star\in(0,1)} =& \frac{ru(2c(\gamma r+1)+(\lambda-r\phi)^2)-4c(\gamma r+1)(\lambda-r\phi)}{(2c(\gamma r+1)-(\lambda-r\phi)^2)^2}.\end{aligned}$$

The sign of the effect of λ on the ad price is the same as the sign of

$$\underbrace{(\lambda-r\phi)(r(2u+\phi)-\lambda)}_{(+)}\ \underbrace{-2c(\gamma r+1)}_{(-)}.$$

Two opposite effects are at play. First, an increase in the marginal reputation loss for advertisers, λ, has a positive effect on the marginal gains of the platform from raising content moderation. Therefore, the platform tends to increase the price. Second, a negative effect is associated with the cost of content moderation. Thus, a threshold value of c exists below which raising λ has a positive effect on the advertising price and above which it leads to a lower ad price.

The sign of the effect of λ on the moderation level is the same as the sign of

$$\underbrace{ru(2c(\gamma r+1)+(\lambda-r\phi)^2)}_{(+)}\ \underbrace{-4c(\gamma r+1)(\lambda-r\phi)}_{(-)}.$$

Also, in this case, there are two opposite effects. Firstly, a positive effect is associated with the gains from moderation (and users' participation). Secondly, there is a negative effect associated with the moderation cost. The net effect depends on the marginal moderation cost, and there

is a critical value of c above which the effect is positive.

In the Appendix, we show that the critical value of c is the same in the two cases, and we denote it as $\tilde{c} := \frac{(ur)^2}{2(\gamma r+1)}$. The following proposition summarizes the above discussion.

Proposition 2. *Under Assumptions (A1-A5), for any $\phi r < \lambda < \phi r + \frac{2c(\gamma r+1)}{ru}$, a higher λ has the following effects on the equilibrium price and content moderation intensity:*

- *if $c \leq \tilde{c}$, then $p^\star$ is U-shaped in λ and $m^\star$ increases in λ.*
- *if $c > \tilde{c}$, then $m^\star$ is inverted-U shaped in λ and $p^\star$ decreases in λ.*

This proposition can be explained easily with the aid of Figures 1 and 2, which indicate that the equilibrium content moderation policy of the platform is concave in λ, whereas the equilibrium price is convex in λ. If c is sufficiently low, the platform finds adjusting its content moderation policy relatively cheaper because any marginal increase in the intensity of content moderation does not cost much. As a result, the optimal moderation policy is increasing with the reputation loss of advertisers when exposed to unsafe content. However, the price is U-shaped. The intuition is as follows: for a (relatively) low λ, the loss associated with unsafe content is low for advertisers, meaning that the marginal gain from higher moderation is small and insufficient to attract more advertisers. As a result, the platform uses the pricing instrument to attract advertisers, lowering the ad price. For a (relatively) high λ, the loss associated with unsafe content is high for advertisers. This means that marginal gains from moderation are higher for the platform when advertisers benefit more from a safer environment. As a result, the platform can extract a greater surplus by raising the price.

If c is sufficiently high, any marginal increase in content moderation intensity is expensive for the platform. Due to the concavity of $m^\star$ in λ, the platform raises its content moderation intensity only when this loss faced by advertisers is low because it would be too costly to offset advertisers' losses if these are high. Because advertisers' losses increase faster than potential gains from a higher content moderation intensity, the platform finds it optimal to lower its price.

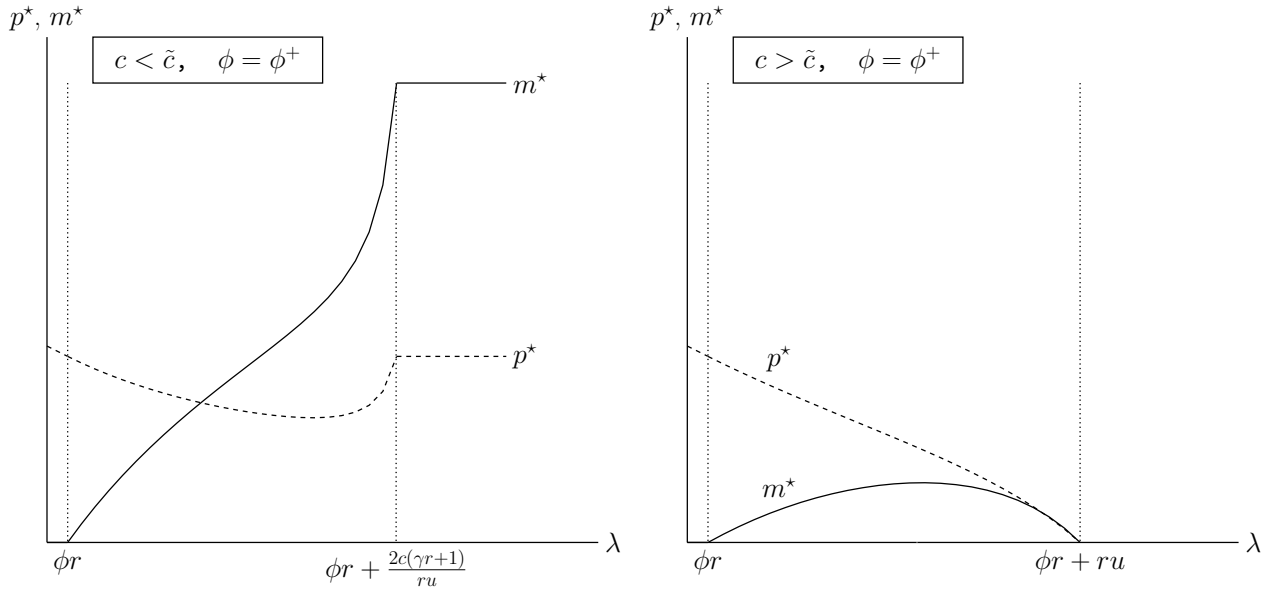

Figure 1: Conflicting tastes for moderation. The impact of a higher brand risk on $p^\star$ and $m^\star$ when c is small, i.e., $c = 0.2$ (left), and large, i.e., $c = 0.5$ (right). Parameter values: $u = 0.9, r = 0.9, \gamma = 0.5, c = 0.5, \phi = 0.05$.

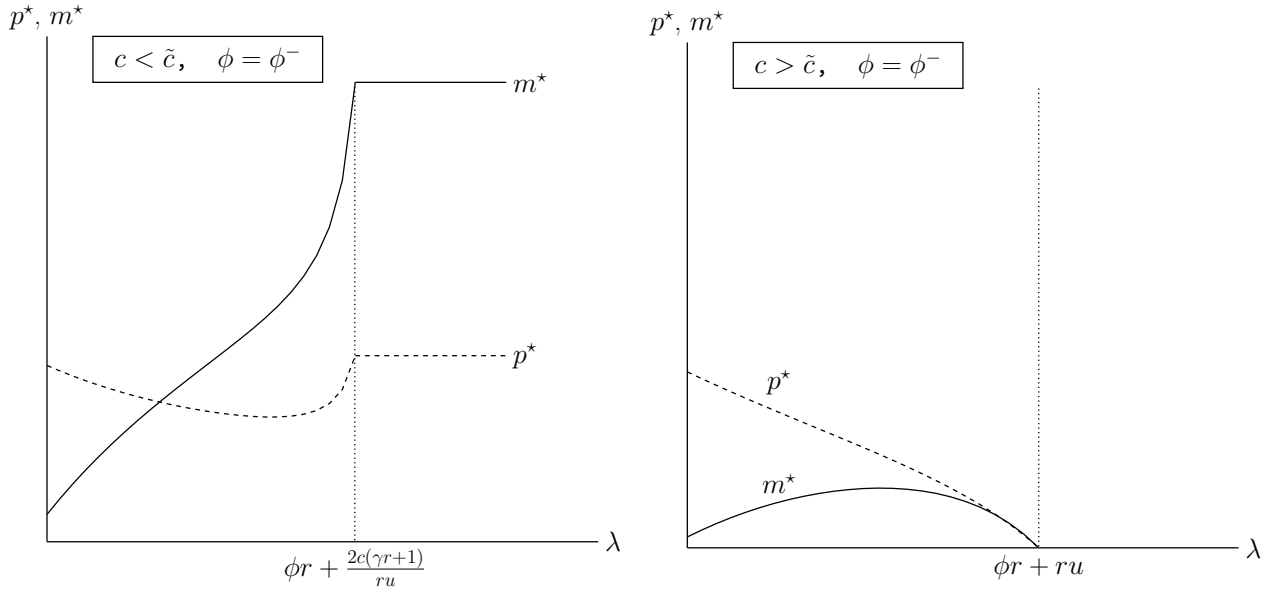

Figure 2: Congruent tastes for moderation. The impact of a higher brand risk on $p^\star$ and $m^\star$ when c is small, i.e., $c = 0.2$ (left), and large, i.e., $c = 0.5$ (right). Parameter values: $u = 0.9, r = 0.9, \gamma = 0.5, \phi = -0.05$.

These results apply regardless of whether users and advertisers have congruent or conflicting preferences. Yet, the nature of their preferences matters for determining the parameter ranges in which the different effects identified are present.

Moreover, we note that network externalities in our framework are particularly important because they generate countervailing incentives for the platform. To see why, suppose users do

not encounter a disutility from the presence of ads, that is, $\gamma = 0$. Our analysis suggests that without advertising nuisance, the platform would have a higher incentive to attract advertisers and, therefore, to increase its content moderation intensity. Importantly, the platform might have an incentive to moderate all unsafe content if moderation is not too costly for the platform.

4.3 Socially desirable level of content moderation

In this section, we compare the level of content moderation the platform provides, with the level that would maximize total welfare. It is important to note that identifying the correct measure of welfare in such a case is not free of concerns, especially when considering that unsafe content in our setting is not necessarily illegal and, as we discussed, may generate utility for some users (i.e., when $\phi > 0$). For this reason, we begin by assuming that the social planner considers only $\min\{0, \phi\}$. In other words, the utility of a user considered by the social planner is the same as the utility of a user if the user finds the unsafe content unsafe (i.e., congruent preferences), whereas it is zero if the user finds the unsafe content as generating utility (i.e., conflicting preferences).

We define the total welfare as follows:

$$W(m, p) = \Pi(m, p) + AS(m, p) + \tilde{U}S(m, p),$$

where AS and $\tilde{U}S$ are advertisers' and users' surplus, respectively, where

$$AS(m, p) = \int_0^{rn-\lambda\theta(m)-p} [rn - \lambda\theta(m) - p - \xi]h(\xi)d\xi, \tag{5}$$

$$\tilde{U}S(m, p) = \int_0^{u+\phi\theta(m)-\gamma a} [u + \min\{0, \phi\}\theta(m) - \gamma a - \omega]f(\omega)d\omega. \tag{6}$$

We assume that welfare is concave in m and define $m^W := \arg\max_m W(m, p)$. For ease of comparison, we also consider a *broader* approach to welfare in which the social planner also accounts for the surplus that the presence of unsafe content can generate to users, i.e., if

$\phi = \phi^+ > 0$. In this case, the user surplus considered by the social planner is

$$US(m,p) = \int_0^{u+\phi\theta(m)-\gamma a} [u + \phi\theta(m) - \gamma a - \omega] f(\omega) d\omega. \tag{7}$$

and total welfare is now defined as $\hat{W}(m,p) = \Pi(m,p) + AS(m,p) + US(m,p)$ where $\hat{m}^W := \arg\max_m \hat{W}(m,p)$.

We modify the timing of the game as follows: In the first stage, the social planner moves first by choosing m^W. In the second stage, the platform chooses p. Finally, users and advertisers make their decisions simultaneously. We relegate the technical analysis to the Appendix of the paper, and we solve the game by backward induction.

Figure 3 (left panel) identifies parameter ranges in which the platform over- or under-moderates unsafe content relative to the level that maximizes total welfare, conditional on $m^\star \in (0,1)$, as a function of the reputation loss λ and users' preferences for unsafe content, ϕ. The dark gray area identifies regions where $m^\star < m^W$, meaning that the platform chooses too little content moderation relative to what the social planner would do. The light gray area identifies regions where $m^\star > m^W$, meaning that the platform chooses too much content moderation relative to what the social planner would do. The white area identifies the region outside our parameters range.

The figure can be better understood starting from $\lambda = 0$, i.e. when advertisers do not suffer from brand risk. In this case, there is no externality to advertisers, and the platform only uses content moderation to control users' participation, whereas it uses the ad price to control the participation of advertisers. In this case, the platform always under-moderates when users and advertisers display congruent preferences. Note that when advertisers' and users' preferences conflict, we end up in a scenario where the market is fully covered on the user side (i.e., (A4-A5) does not hold).

As the reputation loss grows larger (increase in λ), the platform raises its level of content moderation, but whether there is more or less moderation than what the social planner would like depends on ϕ.

Specifically, when users do not care much about the presence of unsafe content, i.e., ϕ is close to zero, the platform over-moderates content. This is because the platform is biased towards advertisers who would suffer significantly from brand safety issues, whereas users would not react much to changes in the level of moderation.

When users' and advertisers' preferences are congruent (i.e., ϕ^- is too negative), instead, the platform is more likely to under-moderate. In this case, the platform internalizes the fact that moderating more would attract additional advertisers, which, in turn, will generate disutility for users due to ad nuisance. This leads to under-moderation relative to what the social planner will choose.

Moving our attention to the area where advertisers and users have conflicting preferences (i.e., $\phi = \phi^+ > 0$), we find that over- and under-moderation can exist. Specifically, since the social planner disregards user utility from consuming unsafe content, it tends to choose a level of moderation m^W that is higher than that chosen by the platform because the latter can attract users by being more lenient. This leads to under-moderation when reputation losses for advertisers are not too large. However, as reputation losses increase, the platform is again biased towards advertisers and resorts to over-moderation relative to the social planner because of the fear of losing too many advertisers.

Figure 3 (right panel) extends the analysis to the broader total welfare definition, where the social planner considers the entire surplus generated by users even when they consume unsafe content. As intuition suggests, for any $\phi = \phi^- < 0$, nothing changes for the social planner and, therefore, $m^W = \hat{m}^W$. Yet, in the area in which users derive benefit from the presence of unsafe content, the platform is more likely to engage in more moderation than the social planner, i.e., $m^\star > \hat{m}^W$.[17] This is because the social planner now puts more weight on the value that users obtain from the content consumed on the platform, even if unsafe, and therefore is more aligned with the platform's incentives. This, in turn, leads to a lower $\hat{m}^W$, whereas the optimal moderation of the platform remains unchanged. As a result, the platform is more likely to over-moderate content relative to the social planner.

[17]Note that the same results would apply when considering the low enough c, i.e., $c = 0.2$, as in Figure 1-2.

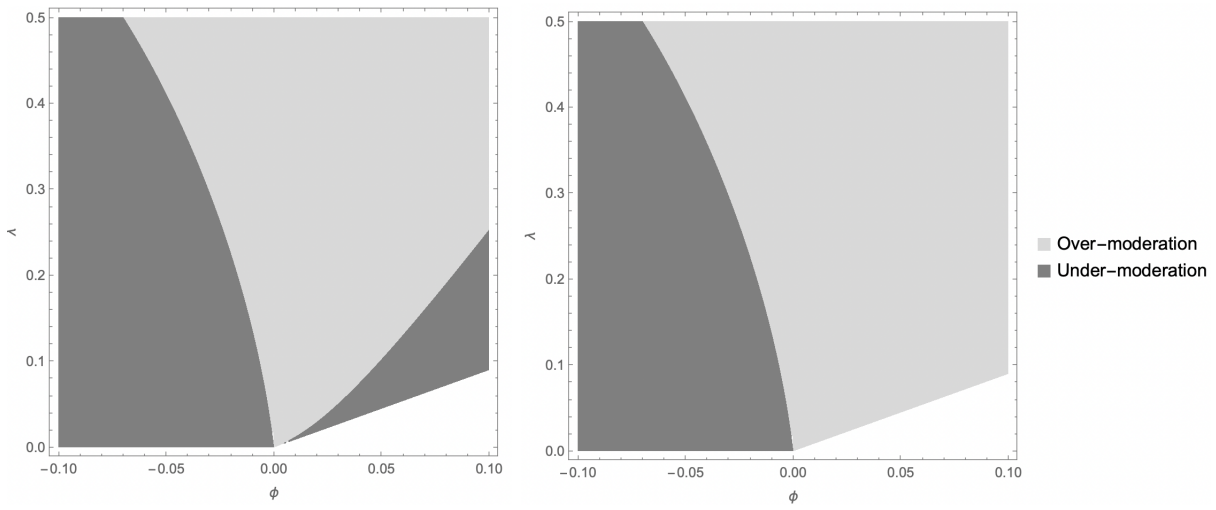

Figure 3: Comparison of the level of content moderation that maximizes the platform's profit, $m^\star$, and the one that maximizes total welfare. In the right panel (resp. left panel), the social planner does (resp. does not) consider positive users' utility from unsafe content (i.e., $\phi > 0$). Parameter values: $u = 0.9, r = 0.9, \gamma = 0.5, c = 0.5$.

5 REGULATION

In the previous section, we addressed the question of whether platforms have appropriate incentives to take down unsafe content and compared their incentives with those of a social planner. We have shown that there is scope for intervention by the social planner since the platform can under-moderate unsafe content, especially when the social planner does not *correct* users' preferences towards unsafe content. In this section, we study how different policies might affect the strategies of online platforms. Firstly, we study the potential unintended effects of inducing online intermediaries to invest more in content moderation through binding mandated content moderation (e.g., Germany's NetzDG).[18] Secondly, we study the effects of taxing online platforms for the externality they generate.

5.1 Mandated content moderation

We begin our analysis by focusing on the effects of mandating platforms to increase their level of moderation above the privately optimal level. This can be the case because the platform

[18]Previous empirical and theoretical studies that have focused on policy intervention have mostly dealt with platform's incentives and copyright-infringing content (Tunca and Wu, 2013; Aguiar et al., 2018; Jain et al., 2020; De Chiara et al., 2021)

is under-investing (as discussed in the previous section) or because the social planner imposes a level of moderation that does not necessarily maximize total welfare, i.e., one-size-fits-all policies that encompass multiple platforms.[19] Indeed, we assume that a regulator or an ad hoc authority obliges online intermediaries to attain a minimum level of content moderation, which we denote as $\hat{m}$. We focus on the scenario in which $m^\star < \hat{m}$ so that the constraint is binding for the platform. Because the platform is constrained in the content moderation decision, the only strategic variable is price. To understand the direction of the strategic response of the platform, we differentiate $p^\star$ and $a(p^\star, m)$ with respect to m. Under Assumptions (A1-A5), we have

$$\frac{dp^\star}{dm} = \frac{\lambda - \phi r}{2} > 0, \qquad \frac{da(m, p^\star)}{dm} = \frac{dp^\star}{dm}\frac{1}{(1+\gamma r)} > 0. \tag{8}$$

for any $\lambda > \phi r$.[20] Therefore, mandated content moderation affords the platform to raise the ad price. Because a higher content moderation intensity leads to a higher utility for advertisers due to the reduced risk associated with unsafe content, the platform finds it optimal to increase its ad price. Importantly, under a uniform distribution, the price increase is less than the increase in the utility of the advertisers. As a result, the platform has more advertising.

Proposition 3. *Under Assumptions (A1-A5), mandating the platform to increase its content moderation level above the privately optimal one, $m^\star \in (0, 1)$, leads to a higher ad price $p^\star$ and a higher number of ads $a(m, p^\star)$ displayed to users.*

This proposition also suggests that, by revealed preferences, advertisers are better off with mandated content moderation. However, this may not necessarily be true for the platform's users. This is because a higher moderation intensity has two effects on users. First, a positive or negative direct effect occurs because a higher moderation can benefit or harm consumers depending on whether they draw utility or suffer from the presence of unsafe content. Second, a negative indirect effect occurs because a higher moderation intensity leads to more ads and

[19] For a discussion, see, e.g., Lefouili and Madio (2022).
[20] If $\lambda < \phi r$, $m^\star = 0$ and there is no strategic response by the platform.

is a greater nuisance to users. Formally,

$$\frac{dn(m,p^\star)}{dm} = \underbrace{-\frac{\phi(2+\gamma r)}{2(1+\gamma r)}}_{\text{direct effect}} \quad \underbrace{-\frac{\lambda\gamma}{2(1+\gamma r)}}_{\text{indirect effect}}. \tag{9}$$

Immediately, a sufficient condition for mandated content moderation to be detrimental to consumers is $\phi = \phi^+ > 0$, which occurs if users enjoy unsafe content. However, if users dislike unsafe content and have congruent preferences, i.e., $\phi = \phi^- < 0$, they face a trade-off between a relatively higher nuisance from ads and a relatively lower presence of unsafe content. The net effect is positive if the increase in the nuisance is less than the gain from a safer platform environment, whereas it is negative otherwise.

We summarize this discussion in the following proposition.

Proposition 4. *Under Assumptions (A1-A5), mandating the platform to increase its content moderation level above the privately optimal one, $m^\star \in (0,1)$, has a positive effect on user participation only if users' aversion to unsafe content is sufficiently strong, i.e.,*

$$-\phi \; > \frac{\lambda\gamma}{(2+\gamma r)}$$

for any $\phi = \phi^- < 0$. In all remaining cases, user participation decreases.

This analysis identifies potential unintended consequences for users when a mandated content moderation policy is imposed. Suppose users and advertisers have congruent preferences, with users suffering significantly from the presence of unsafe content (i.e., $-\phi > \frac{\lambda\gamma}{(2+\gamma r)}$). Then, inducing the platform to raise its content moderation intensity leads to increased participation on both sides of the market. By revealed preferences, advertisers' and users' surplus increases,[21] but the platform is weakly worse off because it is forced to choose a sub-optimal content moderation policy. Suppose now that users and advertisers have conflicting preferences or that users suffer only slightly from the presence of unsafe content (i.e., $-\phi < \frac{\lambda\gamma}{(2+\gamma r)}$). In this case, there is a trade-off for policymakers on the impact of mandated content moderation on surplus

[21] Restricting our attention to a mandated content moderation that raises m in the neighborhood of $m^\star$, such that $\frac{\partial \Pi(m,p^\star)}{\partial m}|_{m=m^\star} = 0$, is certainly socially desirable.

reallocation across sides. A higher content moderation intensity would lead to more advertisers and fewer users, negatively affecting the platform's profit.

The analysis in this section identifies unintended (negative) consequences for users when a mandated content moderation policy is imposed. To shed light on possible interventions that a regulator, or more generally, lawmakers, can make, we consider two polar cases in which the platform is obliged either to moderate *all unsafe content* or not to engage in moderation at (unless the material is manifestly unlawful).[22] Using previous results, we state the following.

Proposition 5. *Suppose* $\hat{m}$ *is either* 0 *or* 1

(i) $\hat{m} = 1$ *is preferred by users and advertisers over* $\hat{m} = 0$ *in the presence of congruent tastes for moderation only if* ϕ *is sufficiently negative. In all other cases, users always prefer* $\hat{m} = 0$*, whereas advertisers always prefer* $\hat{m} = 1$*.*

(ii) A platform's profit is higher with $\hat{m} = 1$ *only if the (marginal) moderation cost is sufficiently small.*

The proof can intuitively follow from (8). Between the two *extreme* cases, the outright removal of all unsafe content is only desirable for advertisers and users if their preferences are congruent and gains for users resulting from a reduction of unsafe material more than compensate for the higher ad nuisance. Otherwise, users are better off with no moderation, which generates brand safety issues for advertisers.

Comparing the profit of the platform in the two scenarios, we observe that

$$\Pi(\hat{m} = 1) - \Pi(\hat{m} = 0) \geq 0 \qquad \text{if} \qquad c \leq \frac{(\lambda - \phi r)(2ru - (\lambda - \phi r))}{2(1 + \gamma r)}$$

Interestingly, there are conditions for which the platform finds it optimal to remove all unsafe content. Specifically, the platform is better off completely removing unsafe content if more moderation triggers a demand expansion on the user side that largely offsets the high cost of

[22] For example, $\hat{m} = 0$ is consistent with a radical form of freedom of speech and any form of automatic monitoring is prohibited.

moderation. In all other cases, the removal of all unsafe content could adversely impact the total surplus in the economy.

5.2 Taxing Digital Platforms

Taxing digital platforms for their activity is critical for policymakers. In this section, we discuss the potentially unintended impact of two types of taxes (namely, a tax on ad revenues and a tax on user activity on the platform) on the moderation strategy of a monopolistic platform.

5.2.1 Taxing digital revenues

Suppose that a fixed tax f^a is imposed on ad revenues, such that the net profit of the platform is equal to $\Pi(a, m, p) = a(n, m, p)(p - f^a) - C(m)$.

As intuition suggests, a similar tax directly affects the platform's marginal revenues, reducing the marginal gains of attracting advertisers. To see why, differentiating the profit with respect to m and p and solving the system of equations yields the optimal price and (interior) content moderation policy, respectively:

$$\begin{aligned} p^\star(f_a) &= \frac{c(1+\gamma r)(r(u+\phi)-\lambda) + f^a(c(1+\gamma r) - (\lambda - \phi r)^2)}{2c(1+\gamma r) - (\lambda - \phi r)^2} \\ m^\star(f_a) &= \frac{(\lambda - r\phi)(r(u+\phi) - \lambda - f^a)}{2c(\gamma r + 1) - (\lambda - r\phi)^2}. \end{aligned}$$

where $m^\star \in (0, 1)$ if $\lambda > r\phi$. In order to understand the effect of the tax f^a on equilibrium outcomes, we differentiate $p^\star$ and $m^\star$ with respect to f^a, which yields

$$\begin{aligned} \frac{dp^\star(f_a)}{df^a} &= \frac{c(1+\gamma r) - (\lambda - \phi r)^2}{2c(1+\gamma r) - (\lambda - \phi r)^2}, \\ \frac{dm^\star(f_a)}{df^a} &= -\frac{\lambda - \phi r}{2c(1+\gamma r) - (\lambda - \phi r)^2}. \end{aligned} \tag{10}$$

Because $m^\star$ is an increasing function of the marginal gains from moderation, the higher the tax, the lower the marginal gains from moderation, and consequently, the lower the incentive to moderate unsafe content. Thus, we have $\frac{dm^\star(f_a)}{df^a} < 0$. Notably, this effect is independent of

whether users and advertisers have congruent or conflicting tastes for moderation.

The effect on the ad price is more subtle due to the presence of two opposing forces. Specifically, the sign of $\frac{dp^\star(f_a)}{df^a}$ aligns with the sign of its numerator, i.e.,

$$c(1+\gamma r) - (\lambda - \phi r)^2.$$

The first term in the above equation captures the pass-through of the tax onto the ad price. This first-order effect does drive up the ad price. The second term, instead, captures the complementarity between the ad price and the (reduced) content moderation effort. This second-order effect leads to a lower ad price. Denoting $\tilde{c}_a := \frac{(\lambda-\phi r)^2}{1+\gamma r}$, the prevailing effect depends on the magnitude of the moderation cost. If the moderation cost is low enough, i.e., $c < \tilde{c}_a$, advertisers are granted a price discount to compensate for the high brand risk. However, if the moderation cost is high enough, i.e., $c > \tilde{c}_a$, advertisers pay a higher price when an ad tax is introduced.

Note that reduced content moderation leads to more unsafe content and negatively affects advertisers' participation levels because

$$\frac{da(m^\star, p^\star)}{df^a} = -\frac{c}{2c(1+\gamma r) - (\lambda - \phi r)^2} < 0.$$

To understand the effect of an increase in the tax on user participation on the social media platform, we differentiate $n(m^\star, p^\star)$ with respect to f^a, which yields

$$\frac{dn(m^\star, p^\star)}{df^a} = \frac{c\gamma + \phi(\lambda - \phi r)}{2c(1+\gamma r) - (\lambda - \phi r)^2}.$$

Thus, $\frac{dn(m^\star,p^\star)}{df^a}$ has the same sign as the sign $c\gamma+\phi(\lambda-\phi r)$, which is always positive if $\phi = \phi^+ > 0$ as $\lambda - \phi r > 0$ to ensure that $m^\star \in (0,1)$. On the other hand, if $\phi = \phi^- < 0$, the effect is positive (respectively, negative) if $c > (<) - \frac{\phi(\lambda-\phi r)}{\gamma}$. This result implies that users are likelier to be better off unless they derive a large utility from the moderation of unsafe content. In the latter case, the gains from a lower ad nuisance are fully offset by the distress of being exposed to unsafe content.

We summarize this result in the following proposition.

Proposition 6. *Suppose a tax f^a is levied based on the platform's revenues. An increase in the tax always leads to less content moderation, whereas it leads to a higher advertising price if $c > \tilde{c}_a$ and to a lower advertising price if $c < \tilde{c}_a$. Moreover, advertisers' participation decreases with the tax.*

5.2.2 Taxing platform activity

An alternative form of taxation concerns data collection (Collin and Colin, 2013). For tractability, suppose users are homogeneous in their activity only, so taxing data collection is equivalent to imposing a tax per user. Denote such a tax by f^n, such that the net profit of the platform is equal to $\Pi(a, m, p) = a(n, m, p)p - C(m) - n(a, m)f^n$.

We focus on the most interesting case in which the optimal content moderation policy is interior, i.e., $m^\star \in (0, 1)$. The optimal price and (interior) content moderation policy are

$$\begin{aligned} p^\star(f^n) &= \frac{c(1+\gamma r)(ru - (\lambda - \phi r)) + f^n((\gamma\lambda + \phi)(\lambda - \phi r) - c\gamma(1+\gamma r))}{2c(1+\gamma r) - (\lambda - \phi r)^2} \\ m^\star(f^n) &= \frac{(\lambda - \phi r)(ru - (\lambda - \phi r)) + f^n(\gamma(\lambda + \phi r) + 2\phi)}{2c(1+\gamma r) - (\lambda - \phi r)^2}. \end{aligned} \tag{11}$$

Differentiating $p^\star(f^n)$ and $m^\star(f^n)$ with respect to f^n yields

$$\begin{aligned} \frac{dp^\star(f^n)}{df^n} &= \frac{(\gamma\lambda + \phi)(\lambda - \phi r) - c\gamma(1+\gamma r)}{2c(1+\gamma r) - (\lambda - \phi r)^2} \\ \frac{dm^\star(f^n)}{df^n} &= \frac{\gamma(\lambda + \phi r) + 2\phi}{2c(1+\gamma r) - (\lambda - \phi r)^2}. \end{aligned}$$

Unlike a tax levied on a platform's revenues, introducing a tax proportional to the platform's user base can have either a positive or negative effect on the platform's content moderation policy. Specifically, $\frac{dm^\star(f^n)}{df^n}$ shares the same sign as its numerator. A sufficient condition for $\frac{dm^\star(f^n)}{df^n} > 0$ is that $\phi = \phi^+ > 0$. This implies that a tax based on user activity makes attracting users more expensive, potentially biasing the platform's strategy toward advertisers. Consequently, the larger the tax, the greater the distortion introduced, and the stronger the

incentive for the platform to invest in content moderation when users benefit from the presence of unsafe content.

If $\phi = \phi^- < 0$, users suffer from the presence of unsafe content, two opposing effects exist. On the one hand, attracting more users becomes more costly for the platform, leading to increased content moderation efforts to attract advertisers and, consequently, increasing the nuisance to users. On the other hand, the platform has incentives to lower content moderation efforts since this directly harms users. This is because more stringent content moderation would attract a large mass of users, thereby increasing the user base subject to a tax. As a result, $\frac{dm^\star(f^n)}{df^n} > 0$ only if $\gamma(\lambda + \phi r) + 2\phi > 0$ and the opposite otherwise.

Turning on the ad price, the tax pass-through is not always fully present at equilibrium. With a lower ad price, more advertisers can join the platform. Because users face a higher nuisance, their participation decreases, as does the negative effect of the tax on the platform's profits. Yet, content moderation leads to a higher advertisers' willingness to pay, meaning the platform could set a higher ad price. Depending on the prevailing effect, which is linked to the size of the moderation cost, the ad price decreases for high moderation costs and increases otherwise. It follows that users and advertisers can be better or worse off if such a tax is imposed depending on the usual trade-off between nuisance from (more or less) ads and user preferences for more or less moderation. Specifically, the sign of $\frac{dp^\star(f^n)}{df^n}$ is the same as the sign of its numerator, i.e., positive, that is if $c < \frac{(\gamma\lambda+\phi)(\lambda-\phi r)}{\gamma(1+\gamma r)} := \tilde{c_n}$.

Proposition 7. *Suppose a tax f^n is levied based on the platform's users. An increase in the tax always leads to more content moderation unless $\gamma(\lambda + \phi r) + 2\phi < 0$ for any $\phi = \phi^- < 0$, whereas it leads to a higher (respectively, lower) ad price if $c < (>)\tilde{c}_n$.*

To understand the effect of the tax on advertisers' demand and, by revealed preferences, surplus, let us differentiate $a(m^\star, p^\star)$ with respect to f^n, which yields

$$\frac{da(m^\star, p^\star)}{df^n} = \frac{c\gamma + \phi(\lambda - \phi r)}{2c(1 + \gamma r) - (\lambda - \phi r)^2}.$$

A sufficient condition for $\frac{da(m^\star,p^\star)}{df^n} > 0$ is that $\phi = \phi^+ > 0$ so that users and advertisers have

conflicting preferences. If, on the other hand, preferences are congruent (i.e., $\phi = \phi^- < 0$), there are two opposite effects, and $\frac{da(m^\star,p^\star)}{df^n} > 0$ if $c\gamma + \phi(\lambda - \phi r) > 0$ and $\frac{da(m^\star,p^\star)}{df^n} < 0$ if $c\gamma + \phi(\lambda - \phi r) < 0$ because the platform is biased towards one side of the market, and content moderation policies can directly attract both advertisers and users.

On the user side, instead, the effect of an increase in the tax f^n has the following effect:

$$\frac{dn(m^\star, p^\star)}{df^n} = -\frac{c\gamma^2 + 2\phi(\gamma\lambda + \phi)}{2c(1+\gamma r) - (\lambda - \phi r)^2}$$

A sufficient condition for $\frac{dn(m^\star,p^\star)}{df^n} < 0$ is that preferences are conflicting, i.e., $\phi = \phi^+ > 0$. If, on the other hand, preferences are congruent (i.e., $\phi = \phi^- < 0$), there are two opposite effects, and $\frac{dn(m^\star,p^\star)}{df^n} < 0$ if $c\gamma^2 + 2\phi(\gamma\lambda + \phi) > 0$ and the opposite otherwise. This result suggests that introducing a tax based on user activity on the platform has a non-trivial effect on content moderation policy and the participation of advertisers and users in the platform's activities.

6 EXTENSIONS

In this section, we extend our baseline analysis. First, we study the case of (symmetric) platform competition and explore the potentially unintended effects of fiercer competition on moderation strategies. Second, we relax the assumption of reputational loss being independent of the number of users on the platform and examine how this assumption impacts the main results from the baseline model.

6.1 Platform competition

Social media platforms compete with one another by providing differentiated services. For example, Instagram competes for user attention against TikTok and Snapchat. In this section, we extend our analysis to competing platforms to study how, in a simplified setting with symmetric social media platforms, content moderation policies are chosen, and these are affected by the intensity of competition between platforms. We build on a standard Hotelling model,

with platforms at the endpoints (0 and 1) of a line of unit length. We focus on a competitive bottleneck setting: users only join one platform, whereas advertisers multi-home on both platforms. Platform $i = 1, 2$ maximizes profits by choosing its content moderation policy m_i and the ad price p_i. Profits are $\Pi_i(a_i, m_i, p_i) = a_i p_i - C(m_i)$, where $C(m_i) = cm_i^2/2$

Throughout the analysis, we maintain the same assumptions as in the baseline model and adapt those in (A1-A5) to the current context. Specifically, we assume advertisers are distributed uniformly according to their outside options in $[0, 1]$. For tractability, we assume full market coverage on the user side, and we capture heterogeneity among users in their preference for either platform. We, therefore, assume that users are distributed uniformly on the Hotelling line and their location is indexed by y.[23] Therefore, the utility of a user located at y from joining platform i is given by $U_i(a_i, m_i, p_i) = u + \phi\theta(m_i) - \gamma a_i + T_i(\tau, y)$, with $T_1(\tau, y) = -\frac{\tau y}{2}$ and $T_2(\tau, y) = \frac{\tau y}{2}$, $y \in [\underline{y}, \overline{y}]$ and $\overline{y} = -\underline{y}$, the user relative preference for platform 2. This means that users are distributed symmetrically around zero.

We assume that expectations on the market participation level are fulfilled at equilibrium and focus on a symmetrical equilibrium. We relegate the technical details to the Appendix so that we can express the number of ads and users on each platform as a sole function of each platform's moderation policy and price, i.e., $a_i(m_i, m_j, p_i, p_j)$ and $n_i(m_i, m_j, p_i, p_j)$. As in the baseline model, $\frac{da_i}{dm_i}$ is critical in shaping platform i incentives to engage in moderation and the optimal pricing strategy:

$$\frac{da_i(m_i, m_j, p_i, p_j)}{dm_i} = \frac{\lambda(2\tau + \gamma r) - r\phi}{2\tau},$$

which can be either positive or negative. Importantly, if $\phi = \phi^- < 0 < 0$, users dislike unsafe content and $\frac{da_i(m_i, m_j, p_i, p_j)}{dm_i} > 0$. Therefore, a higher moderation intensity ensures higher brand safety and user participation, increasing advertisers' participation in platform i (all else being equal). Alternatively, if $\phi = \phi^+ > 0$, users generate positive utility from unsafe content, which conflicts with advertisers' preferences. In this case, the sign of $\frac{da_i(m_i, m_j, p_i, p_j)}{dm_i}$ depends on the trade-off between the brand safety effect, now augmented for the intensity of platform

[23] To ensure full market coverage, we assume that u is large enough.

competition for users (captured by τ), and the eyeball effect. The eyeball effect is positive if, for a given nuisance, τ is sufficiently large whereas it is negative if τ is small enough. Indeed, if competition for user attention grows fiercer, the user transportation cost τ would decrease, and users who enjoy unsafe content would move to the rival platform (for a given rival's moderation policy). Consequently, the number of advertisers that join the platform decreases. The opposite would hold if the competition between social media platforms were softened, meaning when facing an unwanted higher content moderation intensity, users would find it too costly to move to the rival platform. This would create an incentive for advertisers to keep advertising on the platform.

The following lemma presents the equilibrium content moderation policy and prices under the assumption of a uniform distribution of the opportunity costs.

Lemma 2. *Consider social media platform competition. The platform sets the following content moderation policy and price:*

(i) If $\lambda \leq \frac{r\phi}{2\tau+\gamma r}$*:*

$$m_i^\star = 0 \qquad p_i^\star = \frac{(\tau+\gamma r)(r-2\lambda)}{4\tau+3\gamma r}.$$

(ii) If $\frac{r\phi}{2\tau+\gamma r} < \lambda < \frac{2c(4\tau+3\gamma r)+\phi r^2}{r(2\tau+\gamma r)}$*:*

$$m_i^\star = \frac{(\lambda(2\tau+\gamma r)-r\phi)(r-2\lambda)}{2(c(3\gamma r+4\tau)-\lambda(\lambda(2\tau+\gamma r)-r\phi))} \qquad p_i^\star = \frac{c(\tau+\gamma r)(r-2\lambda)}{c(3\gamma r+4\tau)-\lambda(\lambda(2\tau+\gamma r)-r\phi)} \quad (12)$$

(ii) If $\lambda \geq \frac{2c(4\tau+3\gamma r)+\phi r^2}{r(2\tau+\gamma r)}$*:*

$$m_i^\star = 1 \qquad p_i^\star = \frac{(\tau+\gamma r)r}{4\tau+3\gamma r}.$$

Competition for user attention creates incentives for platforms to attract users in a twofold manner: By increasing the ad price, the platform can control the number of ads on the platform; by changing moderation, the platform can control the direct effect of moderation in the two sides of the market. The competition for user attention now exacerbates the negative effect of an ad price on advertisers' demand. In other words, having more ads also induces users

to switch to another platform, which means fewer ads are present on the platform of origin, thereby decreasing the profit of this platform. Formally, this effect arises because the marginal gain from moderation and the incentives to invest in content moderation decrease.

The effect of competition on platforms' strategies. In what follows, we identify how more intense competition between platforms affects the incentives to invest in content moderation. Here, we provide simple comparative statics to understand the effect of a change in the transportation cost. We restrict our attention to $\frac{r\phi}{2\tau+\gamma r} < \lambda < \frac{2c(4\tau+3\gamma r)+\phi r^2}{r(2\tau+\gamma r)}$ that is when $m^\star \in (0,1)$. Differentiating $p_i^\star$ and $m_i^\star$ with respect to τ, we obtain:

$$\begin{aligned}\frac{\partial p_i^\star}{\partial \tau} =& \frac{cr(\lambda(\lambda\gamma+\phi)-c\gamma)(r-2\lambda)}{(c(3\gamma r+4\tau)-\lambda(\lambda(2\tau+\gamma r)-r\phi))^2}\\ \frac{\partial m_i^\star}{\partial \tau} =& \frac{cr(2\phi+\gamma\lambda)(r-2\lambda)}{(c(3\gamma r+4\tau)-\lambda(\lambda(2\tau+\gamma r)-r\phi))^2}\end{aligned}$$

The sign of $\frac{\partial p_i^\star}{\partial \tau}$ is the same as the sign of

$$\underbrace{\lambda(\lambda\gamma+\phi)}_{(-/+)} - \underbrace{c\gamma}_{(+)}$$

which is positive when c is low enough, whereas it is negative if c is high enough. Denoting $\tilde{c}_{comp} := \frac{\lambda(\lambda\gamma+\phi)}{\gamma}$, then $\frac{\partial p_i^\star}{\partial \tau} > 0$ for any $c < \tilde{c}_{comp}$ and $\frac{\partial p_i^\star}{\partial \tau} < 0$ for any $c > \tilde{c}_{comp}$.

Moreover, the sign of $\frac{\partial m_i^\star}{\partial \tau}$ is the same as the sign of

$$\underbrace{2\phi}_{(-/+)} + \underbrace{\gamma\lambda}_{(+)}$$

The first term, capturing users' preferences, is positive (respectively, negative) if users enjoy (respectively, dislike) unsafe content. The second term relates to the interplay between the nuisance cost and the brand risk advertisers face. This captures the positive change in advertisers' participation. A sufficient condition for $\frac{\partial m_i^\star}{\partial \tau} > 0$ is that users have conflicting preferences with advertisers for content moderation, i.e., $\phi = \phi^+ > 0$. In this case, reducing τ (i.e., a fiercer competition) negatively affects the equilibrium content moderation.

Under congruent preferences, i.e., $\phi = \phi^- < 0$, there are opposing forces, and the net effect is determined by the prevailing one. Therefore, a stronger competition for users leads to an increase (respectively, decrease) in content moderation intensity only if $\phi = \phi^- < 0$ is negative enough, that is $\frac{\lambda\gamma}{2} < -\phi$. These results are summarized as follows:

Proposition 8. *If competition between social media platforms becomes fiercer on the user side, the ad price $p_i^\star$ decreases (respectively, increases), if $c \leq (>)\tilde{c}_{comp}$ whereas the content moderation intensity decreases (respectively, increases) if $\frac{\lambda\gamma}{2} > -\phi$*

Proposition 8 shows that when competition for users becomes fiercer, platforms use both the ad price and the content moderation intensity to attract users. The moderation effort directly affects users and advertisers, whereas the ad price indirectly affects users' utility.

Suppose that users generate utility from unsafe content ($\phi = \phi^+ > 0$). When competition is fiercer, and c is relatively high, reducing advertising nuisance to attract users (hence increasing the ad price) appears cheaper than increasing the moderation intensity. If c is relatively low, using the content moderation policy is a relatively cheap instrument to attract users. However, increasing content moderation intensity for a given ad price would attract advertisers and generate advertising nuisance to users. Facing this trade-off, fiercer competition for user attention induces the platform to lower its content moderation intensity.

Suppose now that users dislike unsafe content and have congruent preferences with advertisers (i.e., $\phi = \phi^- < 0$). In this case, increasing content moderation gives users a direct utility from less unsafe content but an indirect disutility from seeing more advertisers attracted by a higher brand safety. Thus, as the competition between platforms increases, the platform balances those two effects and only increases its content moderation if the direct effect from unsafe content dominates the advertising nuisance created by more advertisers. Because of this, the platform is less likely to lower advertising nuisance to attract users, because increasing the moderation policy pleases both users and advertisers.

6.2 Proportional reputational loss

In the baseline model, we considered a scenario in which advertisers' losses are exogenous, not directly linked to users' exposure, as not all users decide to purchase the advertised product or service. However, there may be cases where reputation losses are proportional to the number of users exposed to unsafe content displayed alongside an ad. In what follows, we identify under which conditions our main results qualitatively hold and what are the novel effects that will now take place.

Specifically, let us consider the following setting where, for a given price p, advertisers' payoffs are given by the following one

$$[r - \lambda\theta(m)]n - p.$$

where $r - \lambda\theta(m)$ represents the net value generated by each interaction with users. Note that if $r - \lambda\theta(m) < 0$, advertisers obtain a loss regardless of the advertising price. To keep the analysis interesting, therefore, we assume that $r - \lambda\theta(0) > 0$, meaning that in the context of no moderation at all, it is still possible for advertisers to obtain revenues from user interaction.

We can write the mass of advertisers (once accounting for the feedback loop) as follows:

$$a(m,p) = H\Big(F(u + \phi\theta(m) - \gamma a(m,p))[r - \lambda\theta(m)] - p\Big).$$

Differentiating it with respect to m, and dropping the arguments of f and F for ease of notation, yields

$$\begin{aligned}\frac{\partial H(\cdot)}{\partial m} =& h(\cdot)\Big(- \lambda\theta'(m)F(\cdot) + [r - \lambda\theta(m)]f(\cdot)(\phi\theta'(m) - \gamma\frac{\partial H(\cdot)}{\partial m})\Big) \\ =& \frac{h(\cdot)\theta'(m)}{1 + h(\cdot)\gamma[r - \lambda\theta(m)]f(\cdot)}\Big([r - \lambda\theta(m)]f(\cdot)\phi - \lambda F(\cdot)\Big).\end{aligned}$$

Because $\theta'(m) < 0$, then $\frac{\partial a(m,p)}{\partial m}$ has the opposite sign of that of

$$[r - \lambda\theta(m)]f(\cdot)\phi - \lambda F(\cdot) < 0,$$

which is always the case if $\phi = \phi^- < 0$.[24] We can decompose the effects as follows:

$$\underbrace{rf(\cdot)\phi\theta'(m)}_{\text{eyeball effect}} \quad \underbrace{-\lambda\theta(m)f(\cdot)\phi\theta'(m)}_{\text{eyeball mitigation effect}} \quad \underbrace{-\lambda F(\cdot)\theta'(m)}_{\text{brand safety effect}}.$$

Firstly, there is the brand safety effect (as in the baseline model), which increases advertisers' participation since $\theta'(m) < 0$. Unlike the baseline model, now this effect is proportional to the mass of users joining the platform, i.e., $\lambda F(\cdot)$, and therefore lower than in the baseline model, i.e., $\lambda \times 1$. Secondly, there is the eyeball effect, like in the baseline model, i.e., $rf(\cdot)\phi\theta'(m)$, which can be positive or negative depending on the sign of ϕ. Thirdly, another term amplifies or mitigates the eyeball effect, i.e., $\lambda\theta(m)f(\cdot)\phi\theta'(m)$. This term now captures the interplay of the losses or gains for brands resulting from a certain amount of unsafe content and is positive if $\phi = \phi^+ > 0$ and negative otherwise.

Note that akin to the baseline model, a sufficient condition for the number of advertisers to increase with more content moderation is that $\phi = \phi^- < 0$, which implies that $[r-\lambda\theta(m)]f(\cdot)\phi-\lambda F(\cdot) < 0$ and therefore $\frac{\partial a(m,p)}{\partial m} > 0$ because $\theta'(m) < 0$. On the contrary, necessary and sufficient conditions require $[r - \lambda\theta(m)]f(\cdot)\phi - \lambda F(\cdot) < 0$, that is $\phi[r - \lambda\theta(m)]\frac{f(\cdot)}{\lambda F(\cdot)} < 1$.

Thus, introducing per-advertiser reputation alters the effect of raising content moderation on advertising participation by adding a third effect to the ones initially considered. Our main results hold qualitatively, although they are likely to be mitigated or amplified depending on the shape of the distributions.

7 MAIN HIGHLIGHTS AND CONCLUSION

The digital revolution has changed the production of media content. Some of the content, though viral, can be toxic or unsafe. In this article, we study the trade-off faced by advertisers who suffer brand safety issues from the spread of unsafe content and the platform's incentives to

[24] Note that in the baseline model, the condition to be satisfied is

$$rf(\cdot)\phi - \lambda < 0.$$

the first term identifies the eyeball effect, and the second one identifies the brand safety effect.

curb their online presence. In this section, we summarize the main results, identifying critical implications both for managers of brands and media agencies and for policymakers.

Managerial implications. First and foremost, we identify conditions for a platform to invest in costly content moderation. We show that the platform might not invest in content moderation. This case arises only if (i) users strongly prefer unsafe content and (ii) advertisers' losses from the presence of unsafe content are limited. In all other cases, the platform has an incentive to curb unsafe content, although there is a partial content moderation intensity. Our analysis suggests that social media managers should carefully assess how advertisers' and users' preferences align. Although moderating unsafe content has a positive direct brand safety effect for advertisers, it might repel participation by users who like unsafe content. The Tumblr case, as discussed in footnote 5, provides suggestive evidence about the divergence of preferences across sides of the market and how failing to account for them can lead to the destruction of the user base. For the microblogging platform, the change in the moderation policy, motivated by the aim to ensure a brand-safe environment for advertisers, triggered the exit of many content creators and viewers, thus reducing the platform's value.

Second, social media platforms should pay particular attention to factors that can increase (or reduce) brands' sensitivity to unsafe content, as it would affect advertisers' willingness to pay differently. Due to the responsiveness of users to content moderation and advertising, our analysis shows that it may not always be optimal for the platform to raise the intensity of content moderation if advertisers become more sensitive to the presence of unsafe content. This might help explain why advertisers are not always satisfied with social media platforms' moderation strategies and have started boycotting large platforms.

Third, moderation costs also matter and can affect asymmetrically different platforms, further answering why content moderation policies differ. For example, in its moderation report, Facebook states that moderation costs are idiosyncratic to countries, depending on language, culture, and other characteristics.[25] Language barriers are likely to increase moderation costs

[25] A summary of the report can be found on the Transparency page of Facebook. https://transparency.facebook.com/community-standards-enforcement.

because AI moderators might be unable to deal with certain spoken languages or regional dialects.[26]

Other applications. Our setting can offer insights into content moderation policies in other industries. First, consider a (traditional) media outlet, which an editor and an editorial board characterize. These outlets, therefore, have almost full control over the type of content they display. Such a practice differs from platforms that do not control content production. However, even professional content can feature a divergence between the interests of the users and those of the advertisers (see, e.g., Ellman and Germano 2009). For instance, in September 2016, following the online campaign "Stop Funding Hate" related to the presence of disputed content on migrants, advertisers such as The Body Shop, Plusnet, Walkers, and others announced they would stop advertising on *The Daily Mail* and *The Sun*. Such a story fits the trade-off that traditional media outlets may face when producing or reporting potentially controversial content.

An ad-funded news outlet that only produces professional content but is sufficiently attention-grabbing to attract users can represent another example. In this case, the outlet might strategically choose the sensitivity of content to produce to balance user and advertiser preferences. Whereas investments in content moderation might not be required, content production may still be costly. The more professional the content, the higher the cost, and the safer it is for advertisers. However, one may imagine that producing professional content is cheaper than moderating thousands of user-generated content pieces.

Second, consider now that content aggregators host both first-party (i.e., professional content) and user-generated content. An ad-funded content aggregator will choose the share of professional and user-generated content depending on users' and advertisers' preferences. This is akin to the trade-off that the social media platform faces if, for example, one considers that professional content is more costly to produce, advertisers prefer more professional content. In contrast, users might have a preference for or an aversion to the user-generated one.

[26] See BusinessInsider, September 16, 2021 'Facebook's AI moderation reportedly can't interpret many languages, leaving users in some countries more susceptible to harmful posts'.

Finally, we can also apply our framework to TV reality shows, such as the famous *The Big Brother*. Frequently, shows like these are sponsored by advertisers and feature a group of contestants. Although viewers might like houseguest scandals, which keep the reality game alive year after year, advertisers that sponsor the program with their products might not appreciate them. In Italy, in 2018, several different sponsors, including Nintendo, decided to forfeit their partnership with the TV show after it showed bullying in the house.[27] Something similar occurred in France, with advertisers boycotting a TV show because of sensitive content.[28] Indeed, media producers must balance potentially conflicting preferences for borderline content and decide how to moderate what is shown on TV.

Policy implications. Our analysis suggests that online platforms might be too lenient or too strict towards the presence of unsafe content, depending on (i) whether users' and advertisers' preferences are congruent or conflicting, and (ii) whether the social planner accounts for the economic value generated for users from the consumption of unsafe content. This may lead to over- or under-moderation by the platform relative to the content moderation level preferred by the social planner. Our analysis also shows that policymakers might face a trade-off between pleasing advertisers or users when mandating stricter content moderation policies as those provided by the EU Digital Services Act.[29]

Furthermore, we have studied the impact of a digital tax on the platform's incentives to curb unsafe content. We showed that any tax alters platform incentives and induces more or less intensive content moderation depending on how the tax is designed. A tax on user activity would induce more moderation, whereas a tax on ad revenues would induce a lax approach. Moreover, a tax might not necessarily translate into a higher advertising price. Our analysis suggests that particular attention should be placed on the interplay between content moderation policies, pricing strategies, and public policies.

Finally, our analysis identifies a potential trade-off between stimulating competition between

[27] Blitzquotidiamo.com, May 4, 2018, 'Grande Fratello, la grande fuga degli sponsor: niente acqua, shampoo e Nintendo'

[28] LExpress.fr. October 10, 2019

[29] For a discussion on the economic effects of liability for online intermediaries, including social media platforms, see Lefouili and Madio (2022).

social media platforms and guaranteeing a safer social media environment. It shows that a social media platform might lower its content moderation intensity in response to fiercer platform competition for user attention. Our analysis suggests that fiercer competition between (symmetric) platforms may generate a race-to-the-bottom in content moderation efforts, thereby conflicting with the policy goal of ensuring a safe web.

ACKNOWLEDGMENTS

An earlier version of this paper circulated under the title "User-generated Content, Strategic Moderation, and Advertising". We are grateful to the Editor, the Associate Editor, and two anonymous referees for their comments. We also thank Luis Abreu, Malin Arve, Elias Carroni, Alessandro De Chiara, Luca Ferrari, David Henriques, Alexandru Ionescu, Yassine Lefouili, Laurent Linnemer, Christian Peukert, Carlo Reggiani, Michelangelo Rossi, Elia Sartori, Adrian Segura Moreiras, Mark Tremblay, and Patrick Waelbroeck for helpful comments and discussions on previous versions of this paper. The usual disclaimer applies.

References

Abreu, L. and Jeon, D.-S. (2020). Homophily in social media and news polarization. *TSE Working Paper*.

Acemoglu, D., Ozdaglar, A., and Siderius, J. (2021). Misinformation: Strategic sharing, homophily, and endogenous echo chambers. *National Bureau of Economic Research*.

Ada, S., Abou Nabout, N., and Feit, E. M. (2022). Context information can increase revenue in online display advertising auctions: Evidence from a policy change. *Journal of Marketing Research*, 59(5):1040–1058.

Aguiar, L., Claussen, J., and Peukert, C. (2018). Catch me if you can: Effectiveness and consequences of online copyright enforcement. *Information Systems Research*, 29(3):656–678.

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

Andres, R., Rossi, M., and Tremblay, M. J. (2023). Youtube "Adpocalypse"': The Youtubers' journey from ad-based to Patron-based revenues. *ZEW - Centre for European Economic Research Discussion Paper No. 59*.

Andres, R. and Slivko, O. (2021). Content regulation on social media: Evidence from NetzDG. *ZEW-Centre for European Economic Research Discussion Paper*, (21-103).

Beknazar-Yuzbashev, G., Jiménez Durán, R., McCrosky, J., and Stalinski, M. (2022). Toxic content and user engagement on social media: Evidence from a field experiment. *Available at SSRN*.

Belleflamme, P. and Toulemonde, E. (2018). Tax incidence on competing two-sided platforms. *Journal of Public Economic Theory*, 20(1):9–21.

Berman, R. and Katona, Z. (2020). Curation algorithms and filter bubbles in social networks. *Marketing Science*, 39(2):296–316.

Bourreau, M., Caillaud, B., and De Nijs, R. (2018). Taxation of a digital monopoly platform. *Journal of Public Economic Theory*, 20(1):40–51.

Casner, B. (2020). Seller curation in platforms. *International Journal of Industrial Organization*, 72:102659.

Chen, J., Xu, H., and Whinston, A. B. (2011). Moderated online communities and quality of user-generated content. *Journal of management information systems*, 28(2):237–268.

Collin, P. and Colin, N. (2013). Rapport relatif à la fiscalité de l'économie numérique. January 2013.

De Chiara, A., Manna, E., Rubí-Puig, A., and Segura-Moreira, A. (2021). Efficient copyright filters for online hosting platforms. *NET Institute Working Paper*.

de Corniere, A. and Sarvary, M. (2023). Social media and the news: Content bundling and news quality. *Management Science*, 69(1):162–178.

Devaux, R. (2023). Display advertising: How context matters? *Available at SSRN 4352475*.

Ellman, M. and Germano, F. (2009). What do the papers sell? A model of advertising and media bias. *The Economic Journal*, 119(537):680–704.

Jain, T., Hazra, J., and Cheng, T. E. (2020). Illegal content monitoring on social platforms. *Production and Operations Management*, 29(8):1837–1857.

Jeon, D.-S., Lefouili, Y., and Madio, L. (2021). Platform liability and innovation. *NET Institute Working Paper*.

Jiménez Durán, R. (2022). The economics of content moderation: Theory and experimental evidence from hate speech on Twitter. *Available at SSRN*.

Jiménez Durán, R., Müller, K., and Schwarz, C. (2022). The effect of content moderation on online and offline hate: Evidence from Germany's NetzDG. *Available at SSRN 4230296*.

Kind, H. J. and Koethenbuerger, M. (2018). Taxation in digital media markets. *Journal of Public Economic Theory*, 20(1):22–39.

Kind, H. J., Koethenbuerger, M., and Schjelderup, G. (2010). Tax responses in platform industries. *Oxford Economic Papers*, 62(4):764–783.

Kind, H. J., Schjelderup, G., and Stähler, F. (2013). Newspaper differentiation and investments in journalism: The role of tax policy. *Economica*, 80(317):131–148.

Kranton, R. and McAdams, D. (2020). Social networks and the market for news. *Mimeo*.

Lefouili, Y. and Madio, L. (2022). The economics of platform liability. *European Journal of Law and Economics*, 53:319–351.

Liu, Y., Yildirim, P., and Zhang, J. (2022). Implications of revenue models and technology for content moderation strategies. *Marketing Science*, 41(4).

Luca, M. (2015). User-generated content and social media. In *Handbook of Media Economics*, volume 1, pages 563–592. Elsevier.

Mueller-Frank, M., Pai, M. M., Reggiani, C., Saporiti, A., and Simanjuntak, L. (2022). Strategic management of social information. *Mimeo*.

OFCOM (2023). Content moderation in user-to-user online services. *Report*.

Rochet, J.-C. and Tirole, J. (2003). Platform competition in two-sided markets. *Journal of the European Economic Association*, 1(4):990–1029.

Shehu, E., Abou Nabout, N., and Clement, M. (2020). The risk of programmatic advertising: Effects of website quality on advertising effectiveness. *International Journal of Research in Marketing*.

Teh, T.-H. (2022). Platform governance. *American Economic Journal: Microeconomics*, 14(3):213–254.

Tremblay, M. J. (2018). Taxing a platform: Transaction vs. access taxes. *SSRN Working Paper*.

Tunca, T. I. and Wu, Q. (2013). Fighting fire with fire: Commercial piracy and the role of file sharing on copyright protection policy for digital goods. *Information Systems Research*, 24(2):436–453.

Yang, M., Zheng, Z. E., and Mookerjee, V. (2021). The race for online reputation: Implications for platforms, firms, and consumers. *Information Systems Research*, 32(4):1262–1280.

Yildirim, P., Gal-Or, E., and Geylani, T. (2013). User-generated content and bias in news media. *Management Science*, 59(12):2655–2666.

Zhang, K. and Sarvary, M. (2014). Differentiation with user-generated content. *Management Science*, 61(4):898–914.

Appendix

Proof of Lemma 1

The proof immediately follows from the discussion in the main text.

Proof of Proposition 1

The first part of the proof follows immediately from the fact that $m^\star = 0$ if $\frac{\partial \Pi(m,p)}{\partial m}\Big|_{m=0} < 0$. Because $C'(0) = 0$, then $\frac{\partial \Pi(m,p)}{\partial m}\Big|_{m=0} < 0$ if $\frac{dH(.)}{dm}\Big|_{m=0} < 0$, which is the case if $\lambda < rf(\cdot)\phi$.

The second part of the proof determines $m^\star \in (0,1)$ as the solution to

$$\frac{\partial \Pi(m^\star, p)}{\partial m} = 0 : \frac{dH(\cdot)}{dm} - C'(m^\star) = 0.$$

This completes the proof.

Proof of Corollary 1

Under (A1-A5), we write the participation level on the two sides of the market as

$$a(n,m,p) = rn(a,m) - \lambda(1-m) - p,$$

$$n(a,m) = v + \phi(1-m) - \gamma a(n,m,p).$$

Solving for fulfilled expectations, we write the participation level on each side of the market as a sole function of $\{m,p\}$ as

$$a(m,p) = \frac{ru - p - (1-m)(\lambda - r\phi)}{1+\gamma r}.$$

Therefore, the platform profit is

$$\Pi(m,p) = a(m,p)p - C(m) = p \times \frac{ru - p - (1-m)(\lambda - r\phi)}{1+\gamma r} - \frac{cm^2}{2}, \tag{13}$$

which is concave in both arguments under (A2) as

$$\frac{\partial^2 \Pi(m,p)}{\partial p^2}\frac{\partial \Pi(m,p)}{\partial^2 m^2} - \left(\frac{\partial^2 \Pi(m,p)}{\partial m \partial p}\right)^2 = \frac{2c(\gamma r + 1) - (\lambda - \phi r)^2}{(r\gamma + 1)^2} > 0.$$

Moreover, $\frac{\partial^2 \Pi(m,p)}{\partial p^2} = -\frac{2}{\gamma r + 1} < 0$ and $\frac{\partial^2 \Pi(m,p)}{\partial m^2} = -c < 0$.

Differentiating the platform's profit with respect to p and m yields

$$\begin{aligned}\frac{\partial \Pi(m,p)}{\partial m} =& \frac{p(\lambda - r\phi)}{\gamma r + 1} - cm = 0 \\ \frac{\partial \Pi(m,p)}{\partial p} =& \frac{ru - (1-m)(\lambda - r\phi) - p}{\gamma r + 1} - \frac{p}{\gamma r + 1} = 0.\end{aligned}$$

Solving simultaneously, we obtain

$$m^\star = \frac{(\lambda - r\phi)(r(u+\phi) - \lambda)}{2c(\gamma r + 1) - (\lambda - r\phi)^2} \in (0,1) \qquad p^\star|_{m^\star \in (0,1)} = \frac{c(\gamma r + 1)(r(u+\phi) - \lambda)}{2c(\gamma r + 1) - (\lambda - r\phi)^2},$$

Note that the first-order condition with respect to m is negative at $m = 0$ if $\lambda < r\phi$, which implies that $m^\star = 1$. In this case, the optimal price is

$$p^\star|_{m=0} = \frac{r(u+\phi) - \lambda}{2}.$$

Moreover, $m = 1$ occurs if $\lambda \geq \phi r + \frac{2c(\gamma r+1)}{ru}$. In this case, $m^\star = 1$ and the optimal price is

$$p^\star|_{m=1} = \frac{ru}{2}.$$

Proof of Proposition 2

In what follows, we study the impact of λ on the equilibrium outcomes. Differentiating $p^\star$ and $m^\star$ with respect to λ, in the parameter ranges in which $m^\star \in (0,1)$ yields the following:

$$\frac{\partial p^\star}{\partial \lambda} = -\frac{c(\gamma r + 1)(2c(\gamma r + 1) + (\lambda - r\phi)^2 - (\lambda - r\phi)2ru)}{(2c(\gamma r + 1) - (\lambda - r\phi)^2)^2}$$

$$\frac{\partial m^\star}{\partial \lambda} = \frac{ru(2c(\gamma r + 1) + (\lambda - r\phi)^2) - 4c(\gamma r + 1)(\lambda - r\phi)}{(2c(\gamma r + 1) - (\lambda - r\phi)^2)^2}$$

Note that $\frac{\partial p^\star}{\partial \lambda} = 0$, which is the case for $\lambda = \lambda_1$ and $\lambda = \lambda_2$, with

$$\lambda_1 = r(u+\phi) - \sqrt{(ru)^2 - 2c(\gamma r + 1)}$$

$$\lambda_2 = r(u+\phi) + \sqrt{(ru)^2 - 2c(\gamma r + 1)}$$

However, only λ_1 is feasible, which we denote as λ_p. Note that λ_p only exists if $c < \frac{(ur)^2}{2(\gamma r+1)}$. Moreover, as $\frac{\partial^2 p^\star}{\partial \lambda^2}|_{\lambda=\lambda_p} > 0$ meaning that $p^\star$ is convex in λ. If $c > \frac{(ur)^2}{2(\gamma r+1)}$, instead, then $p^\star$ always decreasing in λ $\frac{\partial p^\star}{\partial \lambda} < 0$.

Note that $\frac{\partial m^\star}{\partial \lambda} = 0$ which is the case for

$$\lambda_1 = \frac{\phi u r^2 + 2c(\gamma r + 1) - \sqrt{2}\sqrt{c(\gamma r + 1)(2c(\gamma r + 1) - (ru)^2)}}{ru},$$

$$\lambda_2 = \frac{\phi u r^2 + 2c(\gamma r+1) + \sqrt{2}\sqrt{c(\gamma r+1)(2c(\gamma r+1) - (ru)^2))}}{ru}.$$

However, only λ_1 satisfies our constrains. Denoting it as λ_m, we find that it exists only if $c > \frac{(ur)^2}{2(\gamma r+1)}$. Moreover, as $\frac{\partial^2 m^\star}{\partial \lambda^2}|_{\lambda=\lambda_m} < 0$, $m^\star$ is concave in λ. If $c < \frac{(ur)^2}{2(\gamma r+1)}$, we find $m^\star$ is always increasing in λ.

Proof of Section 4.3

We start solving the game by backward induction. Since the last step of the game, where users and advertisers decide to join the platform, the analysis is the same as in the baseline model. In the second stage, the platform maximizes its profit in (13) by choosing p, which yields

$$p(m) = \frac{1}{2}\Big(r(u + \phi(1-m) - \lambda(1-m))\Big).$$

In the first stage of the game, the social planner chooses m to maximize the total welfare. Plugging $p(m)$ into the advertiser surplus $AS(m,p)$ in (5) yields

$$AS(m, p(m)) = \frac{(\lambda(m-1) + r(u + \phi(1-m)))^2}{8(\gamma r+1)^2}$$

whereas plugging $p(m)$ into the user surplus considered by the social planner, which we denote it as $US(m,p)$ as in (6) yields

$$\tilde{U}S(m, p(m)) = \begin{cases} \frac{1}{2}\left(a^2\gamma^2 - 2a\gamma u - (1-m)^2\phi^2 + u^2\right) & \text{if } \phi = \phi^+ \\ \frac{1}{2}(u - a\gamma + \phi(1-m))^2 & \text{if } \phi = \phi^- \end{cases}$$

and the profit of the platform at $p(m)$ is now

$$\Pi(m, p(m)) = \frac{(r(u + \phi(1-m)) - \lambda(1-m))^2}{4\gamma r + 4} - \frac{cm^2}{2}.$$

The social planner then chooses m to maximize the following

$$W(m, p(m)) = \begin{cases} \Gamma + \frac{((m-1)((3\gamma r+2)\phi - \gamma\lambda) + u(\gamma r+2))(u(\gamma r+2) + (1-m)(\gamma(\lambda + r\phi) + 2\phi)) + (\lambda(m-1) + r(u+\phi(1-m)))^2}{8(\gamma r+1)^2} & \text{if } \phi = \phi^+ \\ \Gamma + \frac{(u(\gamma r+2) + (1-m)(\gamma(\lambda+r\phi)+2\phi))^2 + (\lambda(m-1) + r(u+\phi(1-m)))^2}{8(\gamma r+1)^2} & \text{if } \phi = \phi^- \end{cases}$$

where $\Gamma := \frac{2(\gamma r+1)(\lambda(m-1) + r(u+\phi(1-m)))^2}{8(\gamma r+1)^2} - \frac{cm^2}{2}$.

The socially desirable level of content moderation, denoted as m^W, conditional on being

$$m^W = \begin{cases} \frac{-\gamma^2\lambda^2 - 3\lambda^2 - 2\gamma r^3\phi(u+\phi) + r^2\Omega + r\left(2\gamma^2\lambda\phi - 2\gamma\lambda^2 + 8\gamma\phi^2 + 6\lambda\phi - \left(\gamma^2 - 3\right)\lambda u + 2\gamma u\phi\right) - 2\gamma\lambda u + 4\phi^2}{4c(\gamma r+1)^2 + \phi^2(-2\gamma r^3 + 3(\gamma^2-1)r^2 + 8\gamma r + 4) - \lambda^2(\gamma^2 + 2\gamma r + 3) + 2\lambda r\phi(\gamma^2 + 2\gamma r + 3)} & \text{if } \phi = \phi^+ \\ -\frac{-\gamma^2\lambda^2 - 4\gamma\lambda\phi - 3\lambda^2 - 2\gamma r^3\phi(u+\phi) - r^2\Xi - r\left(2\gamma^2\lambda\phi + 2\gamma\left(\lambda^2 + 2\phi^2\right) - 6\lambda\phi + \left(\gamma^2 - 3\right)\lambda u + 4\gamma u\phi\right) - 2u(\gamma\lambda + 2\phi) - 4\phi^2}{-4c(\gamma r+1)^2 + 2\lambda\phi(2\gamma - 2\gamma r^2 + (\gamma^2-3)r) + \phi^2(2\gamma r^3 + (\gamma^2+3)r^2 + 4\gamma r + 4) + \lambda^2(\gamma^2 + 2\gamma r + 3)} & \text{if } \phi = \phi^- \end{cases}$$

where $\Omega := (\phi\left(3\gamma^2\phi + 4\gamma\lambda - 3\phi\right) + u\left(\gamma^2\phi + 2\gamma\lambda - 3\phi\right))$ and
$\Xi := (\phi\left(\gamma^2\phi - 4\gamma\lambda + 3\phi\right) + u\left(\gamma^2\phi - 2\gamma\lambda + 3\phi\right))$.

Note that to run a comparison with the privately optimal level $m^\star \in (0,1)$ we resort to a

numerical simulation where we consider the following values: $u = 0.9, r = 0.9, \gamma = 5, c = 0.5$.

In turn, focusing on $m^\star(0,1)$ we have the following values

$$m^\star = \frac{(\lambda - 0.9\phi - 0.81)(\lambda - 0.9\phi)}{(\lambda - 0.9\phi)^2 - 1.52}$$

whereas

$$m^W = \begin{cases} \frac{\phi(0.445988 - 1.21651\phi) + 1.\lambda^2 + \lambda(-1.8\phi - 0.495542)}{1.\lambda^2 - 1.8\lambda\phi - 1.21651\phi^2 - 1.01325} & \text{if } \phi = \phi^+ \\ \frac{1.\lambda^2 + \lambda(-1.1012\phi - 0.495542) + \phi(2.20759\phi + 1.98683)}{1.\lambda^2 - 1.1012\lambda\phi + 2.20759\phi^2 - 1.01325} & \text{if } \phi = \phi^- \end{cases}$$

Comparing m^W and $m^\star$ for any $\lambda \in [0, 0.5]$ and $\phi \in [-0.10, 0.10]$ yields results as in Figure 3 (left panel).

Note that results as in the right panel of Figure 3 can be easily obtained from the above equations since the social planner is not constrained in choosing the $\min\{0, \phi\}$ when $\phi = \phi^+$. Therefore, the total welfare is equivalent to the total welfare as when $\phi = \phi^-$, which leads to

$$\hat{m}^W = \frac{1.\lambda^2 + \lambda(-1.1012\phi - 0.495542) + \phi(2.20759\phi + 1.98683)}{1.\lambda^2 - 1.1012\lambda\phi + 2.20759\phi^2 - 1.01325}.$$

Proof of Proposition 3

The proof immediately follows from the discussion in the main text.

Proof of Proposition 4

The proof immediately follows from the discussion in the main text.

Proof of Proposition 5

The first part of the proof follows immediately from the discussion in the text and (8).

The second part of the proof follows. First note that

$$\Pi(\hat{m} = 1) - \Pi(\hat{m} = 0) = \Big[p^\star a(1, p^\star)\Big]|_{m=1} - C(1) - \Big[p^\star a(1, p^\star)\Big]|_{m=0} - C(0).$$

Because $C(0) = 0$ by assumption and $\frac{dp^\star}{dm} > 0$ and $\frac{dp^\star(m,p^\star)}{dm} > 0$, it follows that

$$\Pi(\hat{m} = 1) - \Pi(\hat{m} = 0) > 0 \leftrightarrow C(1) \leq \Big[p^\star a(1, p^\star)\Big]|_{m=1} - \Big[p^\star a(1, p^\star)\Big]|_{m=0}$$

which, with the uniform distribution, implies the following:

$$\Pi(\hat{m} = 1) - \Pi(\hat{m} = 0) > 0 \qquad \text{if} \qquad c < \frac{(\lambda - \phi r)(2ru - (\lambda - \phi r))}{2(1 + \gamma r)}.$$

Proof of Proposition 6

The proof immediately follows from the discussion in the main text.

Proof of Proposition 7

The proof immediately follows from the discussion in the main text.

Proof of Proposition 8

The proof immediately follows from the discussion in the main text.

Proof of Lemma 2

Consider the case of platform competition. Users' demand is denoted by

$$n_i(a_i, a_j, m_i, m_j) = \frac{\tau + (m_j - m_i)\phi + (a_j - a_i)\gamma}{2\tau} \qquad n_j(a_j, a_i, m_j, m_i) = 1 - n_i(a_i, a_j, m_i, m_j). \tag{14}$$

and advertisers' demand is equal to

$$a_i(n_i, m_i, p_i) = rn_i(a_i, a_j, m_i, m_j) - (1 - m_i)\lambda - p_i \qquad a_j(n_j, m_j, p_j) = rn_j(a_j, a_i, m_j, m_i) - (1 - m_j)\lambda - p_j \tag{15}$$

We assume that advertisers and users form expectations that are fulfilled at equilibrium. Solving the associated system of equations yields users' and advertisers' participation as a sole function of (m_i, m_j, p_i, p_j):

$$\begin{aligned} n_i(m_i, m_j, p_i, p_j) =& \frac{(m_j - m_i)(\gamma\lambda + \phi) + \tau + \gamma r + \gamma(p_i - p_j}{2(\tau + \gamma r)} \\ n_j(m_j, m_i, p_j, p_i) =& 1 - n_i(m_i, m_j, p_i, p_j), \end{aligned}$$

$$a_i(m_i, m_j, p_i, p_j) = \frac{(2\tau(m_i - 1) + (\gamma(m_j + m_i - 2))r)\lambda + r(\tau + \phi(m_j - m_i) - \gamma(p_j + p_i)) - 2p_i\tau + \gamma r^2}{2(\tau + \gamma r)}$$

$$a_j(m_j, m_i, p_j, p_i) = \frac{(2\tau(m_j - 1) + (\gamma(m_j + m_i - 2))r)\lambda + r(\tau + \phi(m_i - m_j) - \gamma(p_j + p_i)) - 2p_j\tau + \gamma r^2}{2(\tau + \gamma r)}$$

For ease of exposition, we drop the arguments (m_i, m_j, p_i, p_j). The platforms' profits are

$$\begin{aligned} \Pi_i(\cdot) =& p_i \frac{(2\tau(m_i - 1) + (\gamma(m_j + m_i - 2))r)\lambda + r(\tau + \phi(m_j - m_i) - \gamma(p_j + p_i)) - 2p_i\tau + \gamma r^2}{2(\tau + \gamma r)} - \frac{cm_i^2}{2} \\ \Pi_j(\cdot) =& p_j \frac{(2\tau(m_j - 1) + (\gamma(m_j + m_i - 2))r)\lambda + r(\tau + \phi(m_i - m_j) - \gamma(p_j + p_i)) - 2p_j\tau + \gamma r^2}{2(\tau + \gamma r)} - \frac{cm_j^2}{2} \end{aligned}$$

which are concave in both arguments under the assumption that $c > \frac{((2\tau+\gamma r)\lambda-\phi r)^2}{4(2\tau+\gamma r)(\tau+\gamma r)}$.[30]

From the first-order condition of the platform's profit with respect to p and m we obtain

$$\begin{aligned}
\frac{\partial \Pi_i(\cdot)}{\partial m_i} =& \frac{2p_i\tau\lambda + \gamma p_i r\lambda 2m_i c\tau - p_i\phi r - 2m_i c\gamma r}{2(\gamma r+\tau)} - cm_i = 0 \\
\frac{\partial \Pi_i(\cdot)}{\partial p_i} =& \frac{((2m_i-2)\tau + (m_j+m_i-2)\gamma r)\lambda + (r-4p_i)\tau + \gamma r^2 + ((m_j-m_i)\phi - \gamma p_j - 2\gamma p_i)r}{2(\gamma r+\tau)} = 0 \\
\frac{\partial \Pi_j(\cdot)}{\partial m_j} =& \frac{2p_j\tau\lambda + \gamma p_j r\lambda 2m_j c\tau - p_j\phi r - 2m_j c\gamma r}{2(\gamma r+\tau)} - cm_j = 0 \\
\frac{\partial \Pi_j(\cdot)}{\partial p_j} =& \frac{((2m_j-2)\tau + (m_j+m_i-2)\gamma r)\lambda + (r-4p_j)\tau + \gamma r^2 + ((m_i-m_j)\phi - \gamma p_i - 2\gamma p_j)r}{2(\gamma r+\tau)} = 0.
\end{aligned}$$

Solving simultaneously, we obtain:

$$\begin{aligned}
m_i^\star = m_j^\star &= \frac{(\lambda(2\tau+\gamma r) - r\phi)(r-2\lambda)}{2(c(3\gamma r+4\tau) - \lambda(\lambda(2\tau+\gamma r) - r\phi))} \in (0,1) \\
p_i^\star = p_j^\star &= \frac{c(\tau+\gamma r)(r-2\lambda)}{c(3\gamma r+4\tau) - \lambda(\lambda(2\tau+\gamma r) - r\phi)}.
\end{aligned}$$

Note that $m_i^\star = m_j^\star = 0$ if $\lambda < \frac{r\phi}{2\tau+\gamma r}$. In this case, the optimal price is

$$p_i^\star = p_j^\star = \frac{(r-2\lambda)(\tau+\gamma r)}{4\tau+3\gamma r}$$

Moreover, $m^\star = 1$ if $\lambda \geq \frac{2c(4\tau+3\gamma r)+\phi r^2}{r(2\tau+\gamma r)}$. In this case, the optimal price is

$$p_j^\star = p_j^\star = \frac{r(\tau+\gamma r)}{4\tau+3\gamma r}.$$

[30] Note that

$$\frac{\partial^2 \Pi_i(\cdot)}{\partial p_i^2}\frac{\partial^2 \Pi_i(\cdot)}{\partial^2 m_i^2} - \Big(\frac{\partial^2 \Pi_i(\cdot)}{\partial m_i \partial p_i}\Big)^2 = -\frac{((2\tau+\gamma r)\lambda - \phi r)^2 - 4c(2\tau+\gamma r)(\tau+\gamma r)}{4(\tau+\gamma r)^2} > 0$$

$$\frac{\partial^2 \Pi_j(\cdot)}{\partial p_j^2}\frac{\partial^2 \Pi_j(\cdot)}{\partial^2 m_j^2} - \Big(\frac{\partial^2 \Pi_j}{\partial m_j \partial p_j}\Big)^2 = -\frac{((2\tau+\gamma r)\lambda - \phi r)^2 - 4c(2\tau+\gamma r)(\tau+\gamma r)}{4(\tau+\gamma r)^2} > 0.$$

Moreover, $\frac{\partial^2 \Pi_i(\cdot)}{\partial p_i^2} = \frac{\partial^2 \Pi_j(\cdot)}{\partial p_j^2} = -\frac{2(\gamma r+\tau)}{\gamma r+\tau} < 0$ and $\frac{\partial^2 \Pi_i(\cdot)}{\partial m_i^2} = \frac{\partial^2 \Pi_j(\cdot)}{\partial m_j^2} = -c < 0$.