SELLING CERTIFICATION, CONTENT MODERATION, AND ATTENTION

HESKI BAR-ISAAC[†], RAHUL DEB[≬], AND MATTHEW MITCHELL[‡]

Abstract. Social media platforms moderate content in many ways, balancing the desire of content providers to be seen and trusted with consumers' desire to see and have certified only the content that they value. Content moderation by platforms has come under regulatory scrutiny. We introduce an abstract model of content moderation for sale, where a platform can channel attention in two ways: direct steering that makes content visible to consumers and certification that controls what consumers know about the content before further engagement. The platform optimally price discriminates with both steering and certification, with content from higher willingness-to-pay providers enjoying higher certification and more views. The platform increases profits by cross-subsidizing content from low willingness-to-pay providers that appeals to consumers with higher willingness-to-pay content that does not. This cross-subsidization can also benefit consumers by making content more diverse, suggesting that regulation pushing for accurate certification may be harmful. We identify cases where imperfect certification might be most likely to occur and when forcing higher accuracy would be beneficial.

1. Introduction

The digital environment is overwhelming. There are over a billion websites on the internet, Tiktok hosts billions of videos, more than two trillion posts have been created on Facebook.[1] In such an environment it is little surprise that attention is a key resource and the platforms that control access to such attention can be increasingly sophisticated and profitable in doing so. This has not escaped economists and by now there are surveys on both digital economics (Goldfarb and Tucker, 2019) and social media (Aridor, Jiménez-Durán, Levy, and Song, forthcoming). Meanwhile regulators have become concerned over how platforms, possibly with monopoly power, might steer attention in ways that favor profits over consumer welfare.

Date: March 5, 2025.

[†] heski.bar-isaac@rotman.utoronto.ca; Rotman School of Management, University of Toronto
[≬] rahul.deb@bc.edu; Departments of Economics, Boston College and the University of Toronto
[‡] matthew.mitchell@rotman.utoronto.ca; Rotman School of Management, University of Toronto

We would like to thank Itay Fainmesser, Kinshuk Jerath, Marco Ottaviani, Yossi Spiegel and participants at numerous conferences and seminars for their insightful comments.

[1] These estimates are drawn from https://siteefy.com/how-many-websites-are-there/, https://www.usesignhouse.com/blog/tiktok-stats, and https://www.wordstream.com/blog/ws/2017/11/07/facebook-statistics all accessed on June 7, 2024.

In this paper, we examine how a monopoly platform maximizes profits by governing and charging content providers for two key aspects of attention: how often a post gets viewed (steering) and how it is presented or certified to viewers (certification). We broadly refer to these two channels as content moderation. We show that imperfect certification, where differences that matter to consumers are obscured, can increase content diversity when views are for sale. As a result, consumers can, but need not, benefit from certification for sale relative to enforced perfect certification.

We analyze a model where a platform sells views and certification to content providers interested in consumers' attention. The platform can observe whether a content provider is a good type, valued by consumers, or not; therefore, from the consumer side, the platform has perfect ability (but possibly not incentive) to screen content. Consumers' attention is governed by their intrinsic interest (which makes it ever harder to generate effective views) and by their expectations that any piece of content will be something they value seeing. These beliefs are governed by an understanding of the likelihood of different kinds of content with which they are presented and may depend on how it is presented—which we term certification. In turn these beliefs govern consumers' attention which is what content providers—both good and bad—value. We focus on the application of an online platform selling steering and certification because it corresponds to an important recent development in these markets, and in their regulation. However, in the conclusion section, we point to a wider set of applications that fit our structure.

Content providers want attention: to be seen and, hopefully, trusted so that their content is engaged with. Simply being seen is necessary to command attention and get engagement but not sufficient. In addition to steering content directly, the platform can send an arbitrary message in order to provide consumers with information about quality (certification).[2] This corresponds to everything the consumer sees before deciding whether to further inspect the content. The platform cannot tell the willingness to pay for attention that the content providers have, and they price discriminate by offering different bundles of views and messages.

One might suppose that the ability to control views alone is sufficient to exercise full monopoly power. We show that this is not the case. The ability to vary (and charge for) certification can generate additional platform revenues notably through imperfect certification. That allows for revenue from content that consumers would prefer not to see. In turn, one might suppose that consumers suffer from a platform's ability to sell certification as well as views. We show that there is an economic force acting against this intuition that can overwhelm it. Specifically, consumers

[2] In practice, this may, most obviously, include "checkmarks" for verified status as Twitter began to charge for in November 2022 but could also include position on a page, aspects of display, ancillary information such as "your great aunt Naima likes this post" etc.

might gain from seeing good content whose provider might have relatively little value from receiving this attention. Imperfect certification can, in effect, subsidize views of such content in order to to sell views to bad content and so can improve content diversity.

To see why content diversity is impacted by imperfect certification, start from the case of perfect certification, where the platform is mandated to use its certification technology so that bad types receive no views (one can imagine "bots" are clearly labeled as such) and, so, receive no attention. In our environment, the platform's profit maximization problem of choosing how many views to provide to each content provider and at what price reduces to a classic model of second-degree price discrimination equivalent to the classic analysis of Mussa and Rosen (1978). Since good content providers vary in their valuation for attentive views, those that value attention more, pay more and receive more views. Lower willingness-to-pay get less views, and some positive willingness-to-pay content is excluded entirely.

Now consider the case where there is only one type of certificate available, and its quality is fixed. This is equivalent to all posts receiving the same message but the aggregate share of good content being known. As a result, for each view that any good content provider receives, bad types receive views in some fixed proportion. The platform prefers to sell more good-content views as this allows it to raise revenue from bots; the platform earns revenue from bad types whenever it sells a view to a good content provider. This leads the platform to sell views to good content providers who place relatively little value on gaining attention, in order to show lower quality traffic, offsetting the monopoly distortion in views that arise under perfect certification.

In the fully profit maximizing choice of views and certification, higher willingness-to-pay content providers receive both more views and a message that makes their content more trusted. High enough willingness-go-pay content providers get perfect certification, but lower willingness-to-pay comes with lower certification, but therefore an enhanced return for the platform to show the content of these less willing to pay content providers. As a result, the expansion of content diversity might benefit consumers by making the platform more egalitarian.

To understand why platforms might have changed their approaches to selling attention — notably Twitter's move to charge for "verified status" in November 2022 — we examine comparative statics in the model. In particular, we highlight how a reduction in what platforms can charge for ads can lead to a move away from perfect certification and that cheaper targeting does not affect the quality of certification. In addition, we highlight how the nature of attention affects whether or not platforms engage in imperfect certification—we vary the convexity of attention (corresponding to the extent to which consumers are put off by bad content).

Both steering and certification by platforms have come under regulatory scrutiny more generally. The importance of views is central to algorithmic design around ranking which is under increasing regulatory scrutiny, as in Competition and Markets Authority (2022). The recognition that the presentation of some information might affect the extent to which it is deemed worthy of attention is, of course, at the heart of disclosure regulation (in the context of the kind of social media application that inspires our study, see Mitchell (2021), for example) and central to understanding to understanding and discussion around disclosure and certification. (Dranove and Jin (2010) provides an excellent overview). Content certification for sale has become an issue as well, with the European Commission announcing that they will seek remedies against X for its practice of selling certification through checkmarks.[3]

We show how our result directly impact the policy debate around certification for sale. Our results show that enforcing perfect certification may not benefit consumers. By contrast, we show that when the cost of finding low quality"bots" is low enough relative to the cost of targeting good content, there are sufficient conditions for consumers to benefit from perfect certification being mandated. Imperfect certification, even though it generates increased content diversity, comes with too much bad content in these cases. This highlights the importance of the relative cost of showing different types of content in determining whether or not certification for sale should be regulated, and that there is no simple answer to the welfare impact of enforced perfect certification.

The paper is organized as follows. Next, we review the literature. Then, we introduce the model and construct and simply the mechanism design problem associated with the platform's choices of prices and content moderation (through choosing whether and how much to show different pieces of content and different certification associated with content it shows). Then we consider several benchmarks: an engagement maximizing planner that mirrors consumer welfare, and simple certification with only one or two certificates. The one certificate case includes perfect certification as a special case. Then, we solve the full problem is solved and develop comparative statics. Since it is an important policy concern, we study the comparison to perfect certification in detail, in order to show the trade-offs for consumers in this regulation. Throughout, we illustrate the central forces through an example where attention is a linear function, which allows for an explicit and graphical illustration. Finally, a simple extension that provides intuitive results relating to the role of more damaging bad content and social media addiction. Both can be modelled in a way that makes them impact the planning problem through the way beliefs impact attention in the platform's problem.

[3] See https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761

1.1. *Related Literature*

In considering a platform that sells attention both through certification and more prominent views, we bring together literatures that have considered each of these aspects separately.

1.1.1. *Platform Steering without Certification.*

Our approach is based on solving a second-degree price discrimination problem (in the style of Mussa and Rosen (1978)). Others have considered that the two-sided nature of platforms changes the standard analysis when the platform collects revenue from both sides. Papers include Choi, Jeon, and Kim (2015), Böhme (2016), and Jeon, Kim, and Menicucci (2022).

Choi, Jeon, and Kim (2015) and Böhme (2016) explore how platforms can maximize profits by differentiating prices for different sellers or users, which indirectly steers users. Jeon, Kim, and Menicucci (2022) further expand on this by investigating second-degree price discrimination in monopoly platforms. Focusing on the impact of platform steering on sellers' incentives, Johnson, Rhodes, and Wildenbeest (2023) and Ichihashi and Smolin (2023) argue that steering alters sellers' competitive behavior. By influencing the exposure of sellers' products to consumers, platforms shape the strategic decisions that sellers make. Our model focuses on the social media context, where content is not priced directly by providers to consumers and can be steered to many consumers at low cost. Moreover, the platform sells not only "quantity" in the form of views but also quality (via certification), which affects consumer attention. It is precisely the interaction between these two that is the focus of our analysis.

1.1.2. *Certification without Steering.*

Dranove and Jin (2010) provide a wide-ranging survey of the literature on certification. In this literature, Lizzeri (1999) is an early contribution that shares our conclusion that imperfect certification is optimal as it allows good but not great sellers to charge a higher price.[4] Among more recent papers, perhaps Bouvard and Levy (2018) is the most related in very clearly highlighting that profit-maximizing certification trades off pooling different types of sellers to earn more from low-quality sellers but diluting quality too much alienates consumers (which in our environment corresponds to receiving less attention). Our certification is mixed with direct steering, so that the certification need not do the steering on its own.

1.1.3. *Indirect Steering through Search.*

Direct steering means that the platform in our model must consider the scarcity in the total possible attention (which we capture by a convex cost for the

[4] This literature has developed in several ways. For example, Ali, Haghpanah, Lin, and Siegel (2022) consider uninformed sellers who can conceal that they have been tested. Following the subprime crisis in 2007, a broad literature has considered certification in the credit-rating industry for which useful surveys can be found in White (2010) and Jeon and Lovo (2013).

platform in finding relevant viewers). Consequently, our analysis shares features with the literature that has focused on how platforms sell off scarce slots (including seminal contributions by Edelman, Ostrovsky, and Schwarz (2007), Chen and He (2011), Armstrong and Zhou (2011) and Athey and Ellison (2011), or, more recently, Bar-Isaac and Shelegia (2022) who contrast different sales mechanisms). We share with much of this literature the observation that given the mechanisms through which these positions are sold, consumers draw equilibrium inferences about the quality of offerings associated with their rankings (different certificates, in our work).

Our focus on how profit incentives might lead a platform away from perfect certification and efficiently allocating views is somewhat reminiscent of a literature that examines biased intermediaries and search diversion (De Corniere and Taylor (2014), Hagiu and Jullien (2014), Burguet, Caminal, and Ellman (2015), and De Corniere and Taylor (2019)), though much of this literature is more focused on the consumer search process.

1.1.4. *Content Moderation (Not for Sale).* Content moderation without direct monetary incentives has become an increasingly important area of study as platforms attempt to balance openness with the need to manage harmful or misleading content. Madio and Quinn (2024) study a platform that manages the value of a third party (advertisers) interest in content moderation. (Zou, Wu, and Sarvary, 2025) like this paper highlights a tradeoff between quality and variety but in an environment where entry and content quality are endogenous and affected by a recommendation system that aims to maximize consumer surplus and does not earn revenue from content providers.

Kominers and Shapiro (2024) explore a sender-receiver game where a platform can moderate content, in the sense of manipulating what is seen by the receiver for any message sent. This conforms to our idea of the general description of content moderation, but in a different modeling setting. Acemoglu et al. (2023) study a model where content moderation is about sharing, and how content sharing between consumers might be regulated. Here content moderation is more easily thought of as relating to the content shared between users and not between providers and consumers. Srinivasan (2023) considers a model where a platform allocates views to different kinds of content directly, and highlights a role for the shape of an "attention labor supply function" in an environment where content providers are unsure of the kind of content that they will produce (in contrast to our focus on genuine versus bot content).

Another form of content moderation, that can be considered part of the certification process, is disclosure regulation, where content that is paid must be combined with a message that indicates that this is the case. Inderst and Ottaviani (2012) examine a general model of disclosure regulation. Mitchell (2021) and Fainmesser and Galeotti (2021) model how disclosure regulations might impact relationships between content providers (”influencers”) and consumers (”followers”). Ershov and Mitchell (Forthcoming) provide evidence on the impact of this form of content moderation.[5]

2. Model and preliminaries

We study a price discrimination problem of a platform through which content providers reach consumers. We now describe the model in detail starting with the content providers.

Content Providers. Content *providers* can either be *g*enuine or *b*ots; in the interest of employing more standard terminology, we simply refer to them as *g*ood and *b*ad respectively. There is a continuum of a unit mass of good providers whose private *value* $\theta \in [0, \overline{\theta}] =: \Theta$ is distributed according to F that has a continuous, positive density $f(\cdot) > 0$. θ captures the extent to which a good content provider values engagement. There is an unlimited mass of bad providers who all have the same value for their content being read.[6]

The amount of *engagement* av_g with content is the product of the number $v_g \in \mathbb{R}_+$ of *interested* views that the platform provides a good content provider and the *attention* $a \in [0,1]$ that these viewers pay to the content. The utility of a good content provider with value θ from a given level of engagement av_g is $\theta a v_g$.

Bad providers only value their content being read since consumers never engage with it. A bad content provider that receives attention $a \in [0,1]$ from $v_b \in \mathbb{R}_+$ viewers receives utility av_b.

There are three differences between good and bad providers. First, good providers care about engagement whereas bad providers only care about being read. This distinction is not yet apparent from the payoffs (since they both depend on the product of attention and views) but will become clear below when we define the platforms costs (in essence, it is more costly for the platform to provide interested views to the good providers). Second, there is no heterogeneity in the marginal valuations of the bad providers.[7] Third, there is a limited mass of good providers but we assume (for realism) that there is an unlimited amount of bad content since, in particular, its generation can be automated.

[5] Papers on rules surrounding deceptive sales practices such as Corts (2013), Corts (2014), Glaeser and Ujhelyi (2010), and Rhodes and Wilson (2018)) also highlight the role that some form of rules on messages might play. This also fits our description of content moderation.

[6] Since costs of showing bad type posts will be linear in their quantity, one can equally assume there are many bad types, or merely that there is a fixed quantity but that each bad type’s posts can be spread widely at constant marginal cost.

[7] This assumption is purely for technical convenience since it avoids the complications that arise with multidimensional private information. It is also consistent with the assumption that there are infinitely many bots since the platform can sell to those with highest value.

Targeting of Content and Platform Costs. The platform can distinguish between good and bad providers.[8] However, the platform does not know good providers' valuations for engagement. The platform chooses the number of views to direct to each provider. Directing v_b untargeted views at a bad content provider costs the platform γv_b where $0 < \gamma < \min\{\overline{\theta}, 1\}$.[9] This opportunity cost reflects, for example, that the platform could instead direct advertisements at consumers.

The same opportunity cost is also present when directing views at good providers. However, these providers only value engaged users and this is the source of an additional cost: the platform needs to search for such users of whom there is a smaller pool.[10] The platform faces a cost $\gamma v_g + c(v_g)$ of providing v_g interested views to a good provider, where $c : \mathbb{R}_+ \to \mathbb{R}_+$ is a strictly increasing, strictly convex and differentiable function that satisfies $c(0) = c'(0) = 0$ and $\lim_{v_g \to \infty} c'(v_g) = \infty$. The convexity of c corresponds to the increasing difficulty of targeting the content of a given good provider with interested viewers as the content is shown more times. Therefore, the convexity is at the level of a given content provider.

Consumers. Consumers observe a message m (described in more detail below) and must decide whether to read a post to learn more about the post.[11] Reading requires paying a cost $q \geq 0$, distributed according to a strictly increasing, differentiable cumulative distribution function $A(q)$ after seeing the message. If they do not read, they get zero; if they do, they learn whether the content is good or bad, as well as if they are interested. If it is good content that matches their interest, they engage with the post and get payoff of 1, if they do not, they get a payoff of zero.[12]

Suppose that a consumer has belief μ about the probability that a piece of content will be good and of interest. The expected payoff from reading the post after seeing m is $\mu - q$; therefore the consumer decides whether or not to read based on whether $q \leq \mu$, i.e. with probability $A(\mu)$. We henceforth refer to A as the *attention function*. The attention function reflects an underlying distribution of costs of wasting time on bad posts, compared to the benefits of reading posts that generate engagement. Moreover, since engagement is the consumer's payoff, the engagement maximizing benchmark is a natural consumer welfare standard.

[8] This is consistent with platforms monitoring consumer reactions and other measures of engagement, as well as observing content directly—that is, platforms can distinguish between good and bad providers exactly as users do but need only do so once on behalf of all potential viewers.

[9] This assumption ensures that it is not automatically unprofitable for the platform to serve either type of providers.

[10] Of course, the platform could send out untargeted views in the hopes of randomly finding interested users. Suppose that the unconditional probability of a user being interested in an untargeted post be λ. A sufficient condition that good content is targeted is that the marginal cost of targeting is not less than λ, so it is cheaper to target then to make enough untargeted posts to get the same amount of interest.

[11] The message m can be broadly understood as reflecting whatever the consumer uses to form beliefs about the quality of content before reading it and may include location on a page, information such as how many others "liked" a post, explicit checkmarks, etc.

[12] Note that only average payoffs in each case are relevant. We discuss further costs of bad content in Section **??**

Although the power we give the platform to identify good content may seem strong, notice that one interpretation is that they have only the information that the viewer would have upon reading the post; the platform pays a negligible cost to "read" content (once divided across the many consumers that see the post).

Platform Pricing. The platform price discriminates by offering a (direct) *mechanism* to providers. The mechanism consists of four functions. These correspond to a message or *certificate* assigned to a good provider claiming to be of type θ; the number of targeted views that such a provider who purchases this certificate receives; the number of untargeted bad provider views, for each type θ, that are assigned θ's certificate; and the price that a good provider pays to receive the certificate (the price paid by bad providers is trivial as discussed below and so not incorporated into the design problem). These functions are written as follows:

$$
\begin{aligned}
&M : \Theta \to \mathbb{R}, \\
&V_g : \Theta \to \mathbb{R}_+, \\
&V_b : \Theta \to \mathbb{R}_+, \\
&P : \Theta \to \mathbb{R}.
\end{aligned}
$$

Note that the only private information is the value of the good providers so the mechanism is a function of Θ. A good provider whose value is θ pays $P(\theta)$ to receive a certificate $M(\theta)$ and $v_g(\theta)$ targeted views, at the same time as bad providers receive $v_b(\theta)$ untargeted views. Notice that the general structure of $M(\cdot)$ allows it to stand in for anything related to how content put in front of the consumer is presented. Thus, it can interpreted as an explicit certificate, or the totality of the context for the content (the number of likes, the location on the page etc.).

For every m in the image of M, we use

$$\mu(m) = \frac{\mathbb{E}\left[V_g(\theta) \mid M(\theta) = m\right]}{\mathbb{E}\left[V_g(\theta) + V_b(\theta) \mid M(\theta) = m\right]},$$

to denote the fraction of good views assigned to certificate m or the *quality* of the certificate for short. When both the numerator and denominator are zero in the above fraction, $\mu(m)$ can be chosen arbitrarily.

The platform's mechanism design problem is

$$\max_{V_g,V_b,M,P} \int_{\Theta} \left[P(\theta) + A(\mu(M(\theta)))V_b(\theta) - c(V_g(\theta)) - \gamma(V_g(\theta) + V_b(\theta))\right] f(\theta)d\theta,$$

(1) subject to

$$\theta A(\mu(M(\theta)))V_g(\theta) - P(\theta) \geq \max\{\theta A(\mu(M(\theta')))V_g(\theta') - P(\theta'), 0\} \qquad \text{for all } \theta, \theta' \in \Theta.$$

The objective function is simply the total payments received by the platform net of the costs of providing the views to the content providers. Each good provider type θ pays the platform $P(\theta)$. Bad providers do not have any private information so the platform simply charges them the utility that they receive from being assigned certificate $M(\theta)$ and receiving $V_b(\theta)$ views, which is $A(\mu(M(\theta)))V_b(\theta)$. As described earlier, $A(\mu(M(\theta)))$ is the fraction of consumers who pay attention to content marked with a certificate $M(\theta)$. The constraint captures both the good providers' incentive compatibility and individual rationality constraints.

The case that $A(1) \leq \gamma$ trivially implies that no views will be directed to bad providers since the costs to the platform would then be higher the value of the views. Therefore, we henceforth focus on the more interesting case $A(1) > \gamma$ and specifically, we normalize $A(1) = 1$.

2.1. *Preliminary Analysis and Simplifcation of the Platform's Problem*

Before we begin analyzing the above problem, a few comments are in order. First, observe that because providers are infinitesimal, a misreport by value θ as a value $\theta' \neq \theta$ does not affect the quality $\mu(M(\theta'))$ of the certificate $M(\theta')$ since a single provider cannot change the fraction of good providers. Second, the above problem (1) bears a similarity to the classic work of Mussa and Rosen (1978). The key difference is that the platform is choosing *both* the quality (μ) and quantity (V_g, V_b) of the product, and that these two are related.

We simplify the platform's problem with the following observation.

Lemma 1. *Take any incentive compatible and individually rational mechanism (V_g, V_b, M, P) with associated quality μ. There exists another incentive compatible and individually rational mechanism $(\tilde{V}_g, \tilde{V}_b, \tilde{M}, P)$ such that, for all $\theta \in \Theta$, $\tilde{M}(\theta) = \theta$ and, both the platform and good providers receive the same payoff as from (V_g, V_b, M, P).*

Proof. See Appendix. □

In words, this lemma simply states that it is without loss to assign a separate certificate to each value θ. This is because different certificates can have the same quality. So we can take any mechanism in which different values are assigned to the same certificate and construct a new mechanism in which all values have distinct certificates but we reassign the bad providers' views in a way that the quality is constant across certificates.

This allows us to rewrite the platform's problem such that they are choosing the quality $\mu : \Theta \to [0,1]$ for each value θ instead of the bad provider views V_b since this is pinned down by the equation $\mu(\theta) = V_g(\theta)/[V_g(\theta) + V_b(\theta)]$. The platform thus solves

$$\max_{V_g,\mu,P} \int_\Theta \left[P(\theta) + A(\mu(\theta))V_g(\theta)\frac{1-\mu(\theta)}{\mu(\theta)} - c(V_g(\theta)) - \gamma\frac{V_g(\theta)}{\mu(\theta)} \right] f(\theta)d\theta,$$

subject to

$$\theta A(\mu(\theta))V_g(\theta) - P(\theta) \geq \max\{\theta A(\mu(\theta'))V_g(\theta') - P(\theta'), 0\} \qquad \text{for all } \theta, \theta' \in \Theta.$$

Note that, in the above integrand, a fraction whose numerator and denominator are both zero takes the value zero. We also note that, as written, the above problem permits us to choose $\mu(\theta) > 0$ and $V_g(\theta) = 0$. We do not explicitly prevent this by imposing an additional constraint as, when $V_g(\theta) = 0$, the objective function and the constraints take the same values for any $\mu(\theta) \in [0,1]$. This mild abuse allows us to state results more concisely without changing their economic content.

Now observe that, if we interpret $A(\mu(\theta))V_g(\theta)$ as the "allocation" when value θ is reported, the incentive compatibility constraint is essentially identical to the standard incentive compatibility constraint of (Mussa and Rosen, 1978). Thus, we can use the standard characterization of incentive compatibility to eliminate the price function P from the platform's problem and restate it as

$$\max_{V_g,\mu} \int_\Theta \left[\left(\phi(\theta) + \frac{1-\mu(\theta)}{\mu(\theta)} \right) A(\mu(\theta))V_g(\theta) - c(V_g(\theta)) - \gamma\frac{V_g(\theta)}{\mu(\theta)} \right] f(\theta)d\theta,$$

(2) subject to

$$A(\mu(\theta))V_g(\theta) \geq A(\mu(\theta'))V_g(\theta') \quad \text{for } \theta \geq \theta',\ \theta, \theta' \in \Theta.$$

In the above objective function

$$\phi(\theta) := \theta - \frac{1-F(\theta)}{f(\theta)}$$

is the standard *virtual value*. The constraint captures the fact that incentive compatibility requires the allocation to be nondecreasing. Note that we also eliminated the individual rationality constraint in the standard way by assigning a utility of zero to a good provider of value $\theta = 0$.

In what follows, we assume the distribution F is such that the virtual value ϕ is nondecreasing. This is a standard technical assumption in mechanism design that can be dispensed with but comes at the cost of complication that we view to be orthogonal to our main economic insights. This implies that, if there are functions $V_g^p(\theta)$ and μ^p that are pointwise solutions

$$(V_g^p(\theta), \mu^p(\theta)) \in \operatorname*{argmax}_{v_g,\hat{\mu}} \left[\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu})v_g - c(v_g) - \gamma\frac{v_g}{\hat{\mu}} \right]$$

for all $\theta \in \Theta$ and that satisfy the monotonicity constraint ($A(\mu^p(\cdot))V_g^p(\cdot)$ is nondecreasing), then these are solutions to (2).

3. Benchmarks

In this section, we derive some benchmarks that provide context for the properties of the optimal mechanism. We first solve a "planner's problem" and then derive the properties of the optimal mechanism when the platform is restricted to offering only two certificates.

3.1. *The planner's problem*

Consider a planner who wants to maximize the level of engagement on the platform, taking costs into account.[13] That is, they want to solve

$$\max_{V_g,\mu} \int_{\Theta} \left[A(\mu(\theta))V_g(\theta) - c(V_g(\theta)) - \gamma \frac{V_g(\theta)}{\mu(\theta)} \right] f(\theta) d\theta.$$

The first term in the objective function above is the engagement $A(\mu(\theta))V_g(\theta)$ of users with the good providers (not the utility $\theta A(\mu(\theta))V_g(\theta)$ of good providers) and the remaining terms are the costs associated with directing both targeted and untargeted views to good and providers respectively. Recall that views to bad providers do not generate engagement.

The solution $(\overline{V}_g, \overline{\mu})$ to the above planner's problem is immediate from pointwise optimization and is given by

$$\begin{aligned} \overline{\mu}(\theta) &= 1, \\ \overline{V}_g(\theta) &= c'^{-1}(1-\gamma) \end{aligned}$$

for $\theta \in \Theta$. Recall that we have normalized $A(1) = 1$.

We flag two properties of this solution. Each is intuitive. The first is that, since directing content to bad providers is costly and does not generate engagement, no views are directed to bad providers. In other words, certification for all $\theta \in \Theta$ is perfect with all certificates having quality one. Second, since the planner only wants to maximize engagement, the same number of views are directed to good providers regardless of their value θ. Such egalitarian traffic will not arise from a profit maximizing platform, since the platform will direct more views to providers with higher willingness to pay for those views.

[13] Under the particular class of attention functions $A(\mu) = \mu^{\alpha}$, user welfare is described by engagement (up to a constant of proportionality). See the Appendix for details.

3.2. *Certification with One Message*

In this section, we study a benchmark in which the platform assigns the same message to all types $\theta \in \Theta$ or, in terms of the simplified problem (2), the function μ is a constant function that takes value $\hat{\mu}$. First, this corresponds to a realistic form of content moderation whereby platforms do not distinguish between different kinds of content on the platform (that is all content is presented in the same way) but still choose how many views to allocate to different content providers. The platform may choose to ban bad provider traffic (perfect certification) or not (imperfect certification).[14] Second, the analysis crisply illustrates how imperfect certification leads to an expansion of the set of types that the platform profitably serves. In this sense, imperfect certification allows for greater content diversity—a central theme of our analysis and one that features in the optimal mechanism. We then show that imperfect certification can raise profits relative to perfect certification.

To this end, we derive the traffic for arbitrary quality $\hat{\mu}$ which we then vary to demonstrate the cross-subsidization effect of bad on good traffic, and to consider a hypothetical policy that limits selling of certification (without restricting steering). We view this as the natural benchmark for how the European Commission claims platforms should operate: certification should not be for sale, and certificates should be a clear statement of quality (as, for instance, they claim should be the case on X).[15] They have not taken or suggested any action against platforms that sell traffic in various ways, however.

Fixing a $\hat{\mu}$, the platform's problem (2) boils down to

$$\Pi^s(\hat{\mu}) := \max_{V_g} \left\{ \int \left(\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) V_g(\theta) - \gamma \frac{V_g(\theta)}{\hat{\mu}} - c(V_g(\theta)) \right) f(\theta) d\theta \right\}$$

subject to

$$V_g(\theta) \geq V_g(\theta') \quad \text{for } \theta \geq \theta',\ \theta, \theta' \in \Theta.$$

It is immediate here that the pointwise optimum satisfies the required monotonicity constraint. Thus, the optimum $V_g^s(\theta)$ either takes the value zero or satisfies the first-order condition

$$\frac{\gamma}{\hat{\mu}} + c'(V_g^s(\theta)) = \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}). \tag{3}$$

Thus,

$$V_g^s(\theta) = c'^{-1} \left(\max \left\{ \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}, 0 \right\} \right).$$

[14] This case is also considered in the literature; for example Srinivasan (2023).

[15] https://ec.europa.eu/commission/presscorner/detail/en/ip_24_3761

A useful first benchmark is perfect certification, $\hat{\mu} = 1$, in which only good providers are assigned views. This sort of perfect certification would avoid a policy maker's complaint that messages are "deceiving." So, suppose that perfect certification were enforced, but steering was still for sale.[16] Then, plugging in $\hat{\mu} = 1$, V_g^s is given by $V_g^s(\theta) = c'^{-1}\left(\max\{\phi(\theta) - \gamma, 0\}\right)$.

This solution exactly mirrors Mussa and Rosen (1978) since there are only good content providers; the benefit is the virtual value, and the cost is the sum of γ and the targeting cost. In this solution, the marginal cost of assigning an additional view to a content provider of type θ is equal to the benefit which is precisely the virtual valuation ϕ (that ensures the appropriate information rents accrue to the good providers). Compared to engagement maximization, where all content gets the same number of views, the monopoly platform creates an asymmetry in the views allocated across θ via the virtual valuation ϕ. Those good providers with higher θ who value being seen more are induced to reveal their higher valuation so as to enjoy a higher number of engaged views.

It is useful to compare the first-order condition (3) when certification is perfect ($\hat{\mu} = 1$) and imperfect ($\hat{\mu} < 1$). In the latter case, the costs on the left-hand side of equation (3) are higher: in order to generate an additional (targeted) view for a good provider and ensure quality $\hat{\mu}$ for the certificate, the platform must also provide $\frac{1-\hat{\mu}}{\hat{\mu}}$ untargeted views to bad providers at a cost of $\gamma\frac{1-\hat{\mu}}{\hat{\mu}}$. The benefits on the right side of the equation are also modified. First, the imperfect certificate leads to diminished attention and engagement (captured by $A(\hat{\mu})$) but the platform also earns revenues in proportion to the additional ($\frac{1-\hat{\mu}}{\hat{\mu}}$) views sold to bad providers. It is precisely these revenues that may lead a platform to optimally choose imperfect certification, as we will show below. Moreover, this has implications for content diversity on the platform, as we explore next.

For any certification level $\hat{\mu}$ with the corresponding optimal V_g^s, let

$$\phi_l(\hat{\mu}) = \phi\left(\inf_{\theta\in\Theta}\{V_g^s(\theta) > 0\}\right)$$

denote the lowest type that receives views, or in other words, is shown on the platform. We define $\phi_l(\hat{\mu}) = \overline{\theta}$ when the infimum is taken over the empty set (that is $V_g^s(\theta) = 0$ for all $\theta \in \Theta$), For instance, the lowest type served under perfect certification corresponds to $\phi_l(1) = \gamma + c'(0) = \gamma$. As described above, the potential for revenue from bad providers can lead to imperfect certification; however, serving them leads to diminished attention. If this attenuation of attention is not too severe, the platform will prefer to serve more (rather than fewer) types of good provider. That is, when $A'(1)$ is small enough, imperfect certification can expand the set of types that are served.

[16] This would seem to be consistent with the European Commissions' concern and possible action regarding X, for example.

Proposition 1. *There exists $\delta > 0$ such that, if $A'(1) < \delta$, there exists μ with $\phi_l(\mu) < \phi_l(1)$*

In the engagement maximizing benchmark, all θ (or equivalently, all ϕ) are served. This is typically not the case if the platform serves as a monopolist with perfect certification since $\phi_l(1)$ will typically be interior. Slightly imperfect certification increases content diversity relative to perfect certification by leading more good content providers to enjoy positive views, and therefore, on this dimension, moves closer to the engagement maximizing benchmark, as long as $A'(1)$ is not too big.

Imperfect certification makes content more profitable to show for low θ. But for high θ, it has the downside effect on profits of reducing attention. While Proposition 1 takes the certificate quality ($\hat{\mu}$) as exogenously given to focus on which good content providers to serve, below we connect $A'(1)$ to the platform's optimal choice of certificate quality and its incentive of the platform to certify imperfectly, when both certification and steering are for sale. Indeed, if $A'(1)$ approaches 0, it is almost costless in terms of attention to offer a slightly less than perfect certificate. More generally, lower $A'(1)$ makes the platform more likely to sell imperfect certification. The reason is similar: it makes it easier to cross subsizide good types with bad since the costs in terms of reduced attention and engagement are lower.

Profits for a given level of $\bar{\mu}$ are

$$\Pi^s(\bar{\mu}) := \max_{V_g(\theta)} \int_0^{\overline{\theta}} \left[\left(\phi(\theta) + \frac{1-\bar{\mu}}{\bar{\mu}} \right) A(\bar{\mu}) V_g(\theta) - c(V_g(\theta)) - \gamma \frac{V_g(\theta)}{\bar{\mu}} \right] f(\theta) d\theta$$

Notice that imperfect certification makes traffic more egalitarian. From the first order condition for $V_g(\theta)$ in equation (3), the right hand side increases in θ at rate $\phi'(\theta)A(\bar{\mu})$, so more imperfect certification makes traffic more equal across types. As we discuss below, this force can have a welfare-enhancing effect since the planner's allocation has perfectly egalitarian traffic.

The following provides a sufficient condition for the profit maximizing simple certification to be imperfect:

Proposition 2. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then the optimal single certification $argmax_{\bar{\mu}} \Pi^s(\bar{\mu}) <$ 1*

If $A'(1)$ is not too big, then the gains from imperfect certification outweight the costs of lost engagement. The return to bot traffic, near perfect certification, $1 - \gamma$, provides the sufficient bound on $A'(1)$. Imperfect certification can be profitable.

Linear Attention. The special case of $A(\mu) = \mu$, which corresponds to a uniform distribution of the cost of reading a post, allows a clear illustration of the forces and characterization described throughout the paper.

For certification $\hat{\mu}$ following equation (3), the solution is $V_g = c^{-1}(max\{\phi\hat{\mu} + 1 - \hat{\mu} - \gamma/\hat{\mu}, 0\})$. For perfect certification, therefore, $V_g = max\{c^{-1}(\phi - \gamma), 0\}$.

Figure 1 shows the number of views a good content provider enjoys (on the y-axis) as a function of their type (reflected by the virtual valuation ϕ on the x-axis) and compares perfect certification (depicted by the red line) and imperfect certification with a single certificate of quality $\hat{\mu} = \frac{1}{2}$ (depicted by the blue line).[17]

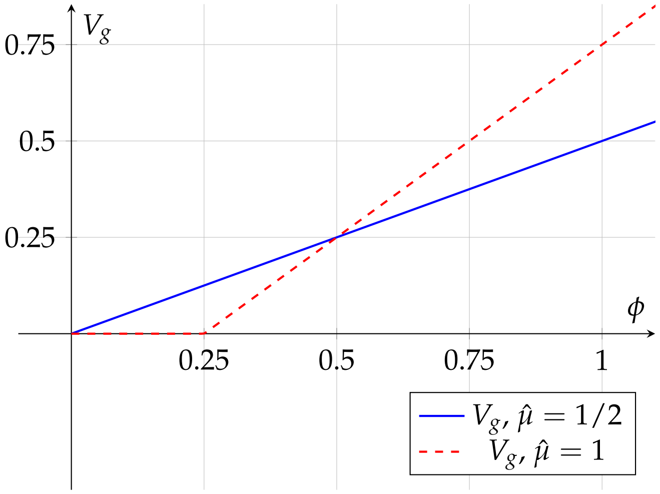


FIGURE 1. Views, Perfect and Imperfect Certification, $\gamma = 1/4$, $c(x) = x^2/2$

The figure highlights the two senses discussed above in which imperfect certification leads to more diversity. First note that a larger number of good content providers receive views under imperfect certification; in particular, there are good providers with lower valuations who receive views when the certificate is imperfect than when it is perfect. Secondly, content is more egalitarian—the number of views is more sensitive to type with perfect certification as can be seen in in the graph through the steeper slope of this relationship. Of course, these notions of improved content diversity do not translate directly into higher welfare, which is captured by engagement. The welfare comparison between perfect certification and the two-certificate case is not immediately obvious from the figure since in shifting from perfect to imperfect certification some (higher) types see lower views, and other types see more views; moreover, all types see less attention conditional on being viewed and in this figure we have not taken a stance on the distribution of types $f(\theta)$. Note that the engagement maximizing benchmark would have identical traffic for all ϕ.

[17] The pictures reflect quadratic costs of targeted posts, but the qualitative features do not rely on this assumption.

3.3. *Two certificates*

The previous discussion of fixed, imperfect certification shows that imperfect certification can expand content. The case with two levels of certification further illustrates how this can improve profits for a platform. The platform can (potentially) benefit from a low level of certification that makes it profitable to serve low valuation good types, cross subsidized with bots, but not necessarily sacrifice engagement on high willingness to pay types.[18] In studying this case, we highlight the sense in which the two instruments, certification and steering, interact. We provide conditions under which the optimal policy has two distinct levels of certification, and many levels of steering; simple certification may be profitable, but is not enough.

Until relatively recently, Instagram and Twitter (now X) had users who were either verified (and their accounts were marked with a check sign) or not (their accounts were unmarked). While they initially did not charge for those providers who obtained a verified status, two certificates is a natural benchmark to study due to this historical precedent. Indeed, when Twitter first started charging for the provision of verified status, they only (in our language) offered two certificates. This section shows how such a structure can improve profits for the platform.

As for the case of a single certificate above, we begin by supposing that the quality associated with the two certificates is exogenously given by $\{\underline{\mu}, \overline{\mu}\}$. We write the two certificate problem as

$$\begin{aligned}\Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin}) := \int_0^{\hat{\theta}^{bin}} & \left[\left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}} \right) A(\underline{\mu}) V_g^{bin}(\theta) - c(V_g^{bin}(\theta)) - \gamma \frac{V_g^{bin}(\theta)}{\underline{\mu}} \right] f(\theta) d\theta \\ & + \int_{\hat{\theta}^{bin}}^{\overline{\theta}} \left[\left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}} \right) A(\overline{\mu}) V_g^{bin}(\theta) - c(V_g^{bin}(\theta)) - \gamma \frac{V_g^{bin}(\theta)}{\overline{\mu}} \right] f(\theta) d\theta\end{aligned}$$

where $\underline{\mu} \leq \overline{\mu}$, since only such an ordering would be incentive compatible. We take $\hat{\theta}^{bin}$ to be interior since we allow $\underline{\mu} \leq \overline{\mu}$.

We provide a complete, though somewhat unaesthetic, characterization in the Appendix. The key features that emerge are that the platform allocates a higher number of targeted views to the higher quality certificate $\overline{\mu}$ and, intuitively, it allocates the higher-quality certificate to higher valuation (high θ) content providers. In this way, the two certificate case echos the finding in the single certificate case that the platform uses the number of views as a price discrimination scheme,

[18] In using certification to soften the incentive constraints for higher types, there is some similarity to (Deneckere and Preston McAfee, 1996). Of course, in our environment the platform earns revenue (from bad providers) in "damaging" the good rather than incurring costs. More substantively, in (Deneckere and Preston McAfee, 1996) consumers are constrained to unit demand, whereas we vary both the number of views and the quality of the certificate leading to somewhat different analysis and effects.

but also uses the differing certificates in this way with higher-value good providers willing to pay both for more views and the greater attention that comes with a higher-quality certificate.

We seek to know when $\underline{\mu} < \overline{\mu}$. That is when two certificateets will be used. To see an example where this is the case, suppose that

$$\phi(\overline{\theta})A'(1) < 1 - \gamma$$

as is sufficient for imperfect simple certification to arise. We will show that the planner cannot simultaneously be optimally choosing some imperfect certification level that suits both the types above and below $\hat{\theta}^{bin}$. The intuition is that the optimal simple certification must trade off the ways in which it deviates from the first order condition for different types; it must be above for some, and below for others. Since incentives are monotone in θ, when this is true for the entire range of θ under simple certification, targetting a level of μ to one type or the other has to improve profits. The sufficient condition guarantees that if $\underline{\mu} = \overline{\mu}$, so certification is simple, it must be interior, and therefore there is room for targetting to improve profits.

Proposition 3. *Suppose $\overline{\theta} \geq 1$, $\phi(\overline{\theta})A'(1) < 1 - \gamma$, and $\phi(\theta)$ is strictly increasing. Then the optimal solution to $max_{\underline{\mu},\overline{\mu},\hat{\theta}^{bin}} \Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin})$ has $\underline{\mu} < \overline{\mu}$.*

This result establishes that two certification levels can dominate one and highlights the intuition that the platform may seek relatively high-quality certificates so as not to dilute earnings from high-value good providers, while using low-quality certificates to use low-quality good providers as a means of earning bot revenue. We next turn to the fully optimal mechanism, where any number of messages can be used and the forces outlined in both the single-certification and two certification cases come to the fore—that is the use of allocating views to price discriminate, coupled with the additional use of different quality certificates to discriminate among different value good providers, while using this diulted quality to earn revenue from bots.

4. Certification and steering for sale

4.1. *The optimal mechanism*

We are now in a position to characterize and analyze the optimal mechanism. The characterization builds on the intuition in Section 3. Relative to the planner's problem, the profit-seeking platform may use imperfect certificates as a means of raising revenue from sales to bots and can more profitably price discriminate between good content providers through offering combinations of both more targeted views and more attention through higher quality certificates. Of course, these considerations interact. The following proposition is a natural generalization of the two certificate benchmark: the pointwise optimum satisfies the required conditions to ensure incentive compatibility and both the views V_g and the quality μ are nondecreasing in θ.

Proposition 4. *There is an optimal mechanism (V_g^*, μ^*) solving the platform's problem (2) where both V_g^*, μ^* are nondecreasing and satisfy*

$$\mu^*(\theta) = \max\left\{\tilde{\mu} \;\middle|\; \tilde{\mu} \in \underset{\hat{\mu}\in[0,1]}{\operatorname{argmax}}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\}\right\} \;>\; 0,$$
$$V_g^*(\theta) = c'^{-1}\left(\max\left\{\left[\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right] A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)}, 0\right\}\right)$$

for all $\theta \in \Theta$.

Recall that incentive compatibility in this setting implies that the optimal mechanism is such that the product $A(\mu^*(\cdot))V_g^*(\cdot)$ is non-decreasing; thus, the main qualitative contribution of the above result is to show that both $\mu^*(\cdot)$ and $V_g^*(\cdot)$ are each individually nondecreasing. This implies that good providers with a higher valuation both receive more traffic and their content is pooled with fewer bad content providers. Certification can be perfect (that is, $\mu^*(\theta) = 1$) for sufficiently high types θ.

The forces that lead to greater content diversity that apply in the simpler benchmarks described above also apply for the optimal mechanism: the presence of low-value good providers can attract more bot revenue; and in addition, the prospect of bot revenue flattens out the relationship between targeted views and good provider valuations leading to more egalitarian content provision than perfect certification, and, in this way, closer to the planner's preference for fully egalitarian content.

To see that the optimal mechanism leads to greater content diversity, in the sense of more good providers being served, it is instructive to compare the optimal (unrestricted) mechanism to a mechanism restricted to offer only two certificates, as explored in Section 3.3. Relative to any pair of binary qualities $0 < \underline{\mu} < \overline{\mu} \leq 1$, it must be the case that

$$\max_{\hat{\mu}\in[0,1]}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\} \geq \max_{\hat{\mu}\in\{\underline{\mu},\overline{\mu}\}}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\}$$

and consequently, the set of types $\{\theta \in \Theta \mid V_g^*(\theta) > 0\} \supseteq \{\theta \in \Theta \mid V_g^{bin}(\theta) > 0\}$ which receive any views at all is a larger set in the optimal mechanism relative to the binary benchmark. We interpret this as greater content diversity that comes from directing traffic towards low value types θ that are unserved under binary certificates. These are types $0 \leq \theta \leq \underline{\theta}$ for which

$$\max_{\hat{\mu}\in[0,1]}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\} > 0$$

where $\underline{\theta}$ satisfies

$$\left(\phi(\underline{\theta}) + \frac{1-\underline{\mu}}{\underline{\mu}}\right) A(\underline{\mu}) - \frac{\gamma}{\underline{\mu}} = 0.$$

Linear Attention. Returning to the case where the attention function is linear, we can use Proposition 4 to characterize the optimal contract explicitly:

Corollary 1. Suppose $A(\mu) = \mu$. Then:

$$\mu(\theta) = \begin{cases} \sqrt{\frac{\gamma}{(1-\phi(\theta))}} & \phi(\theta) \leq 1-\gamma \\ 1 & \phi(\theta) > 1-\gamma \end{cases}$$

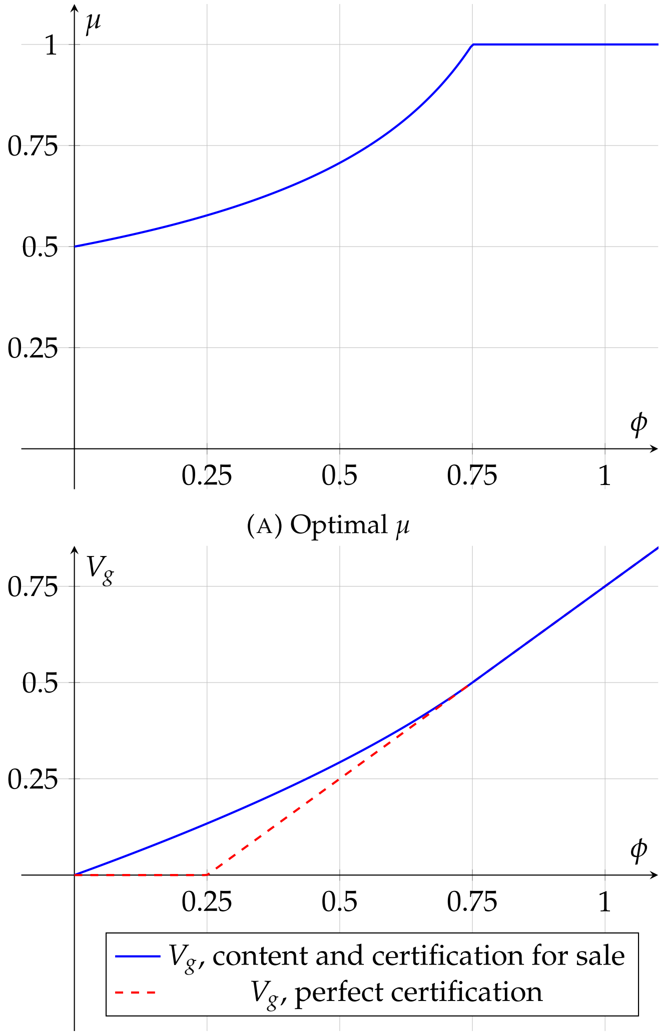
$$V_g(\theta) = \begin{cases} c'^{-1}(\phi(\theta)\mu(\theta) + 1 - \mu(\theta) - \gamma/\mu(\theta)) & \phi(\theta) \geq 1 - \frac{1}{4\gamma} \\ 0 & \phi(\theta) < 1 - \frac{1}{4\gamma} \end{cases}$$

For $\gamma \geq 1/2$, for all θ with non zero views, that is with $V_g(\theta) > 0$ (which requires that $\phi(\theta) < 1 - \frac{1}{4\gamma}$) will have $\mu(\theta) = 1$ (since$\phi(\theta) < 1 - \frac{1}{4\gamma}$ and $\gamma \geq 1/2$ impy that $\phi(\theta) > 1-\gamma$. That is when it is sufficiently costly to supply untargeted views, then the platform will only use perfect certification and the optimal mechanism is perfect certifcation. We therefore focus on $\gamma < 1/2$. In that case the optimal contract uses a variety of levels of certification.

Figure 2 illustrates this. Each panel has the content provider's type on the x-axis, panel (A) showing the quality of the certificate against type, while panel (B) plots the number of views. High-enough value types obtain perfect certification but (as in Corollary 1) differ in the extent to which they receive with higher-valuation providers receiving more views. Lower-value types however receive imperfect certificates with the platform using both instruments—the quality of the certificate and the number of views provided—to price discriminate.

As a result of many certificates, content diversity increases relative to perfect certification, both in terms of types served, and the amount of good posts for types that are served imperfectly when certification is for sale.

In Figure 2(B), we also plot the views allocated under perfect certification. As described above, sufficiently high-value types receive perfect certification and the same number of views as under a perfect certificate. In contrast to the case with only two certificates (illustrated in Figure 1), the sale of lower lower quality certificates to worse types entails a gradual, continuous degradation in certificate quality (rather than a discrete fall) and so, intuitively, for the optimal mechanism there is no distortion to the allocation of the types receiving the perfect certificate. Instead, in the two certificate case, the views of those receiving a perfect certificate must be distorted down in order to ensure that good providers' incentive compatibility constraints are satisfied. Still, the comparison



(A) Optimal μ

(B) Content Diversity

FIGURE 2. Optimal Contract: $A(\mu) = \mu, \gamma = 1/4, c(x) = x^2/2$

between Figure 1 and Figure 2 is suggestive that even with a small number of certificates (as recently introduced on Twitter), a platform may reasonably approximate the optimal solution.

4.2. *Comparative Statics*

The solution described in Proposition 4 allows us to conduct several comparative statics exercises. They help to explain what circumstances do, and do not, lead to imperfect certification. Lower untargeted costs, or a more concave function A, make certification more imperfect. Lower cost of targeting, however, has no impact on certification for a given θ.

Costs of untargeted views. We first examine how the quality of the certificates are affected by the cost γ. Recall, that a natural interpretation for γ, which is the opportunity cost of providing an untargeted view, is lost ad revenue. Thus following result can be understood as stating that as a result of falling ad revenue, good content providers enjoy worse certificates.

Proposition 5. *For all $\theta \in \Theta$, the quality $\mu^*(\theta)$ is nondecreasing in γ.*

To get intuition for the optimal contract and comparative statics on γ, rewrite the first order condition with respect to μ as

$$A(\mu) - \mu^2(\phi + \frac{1-\mu}{\mu})A'(\mu) = \gamma \tag{4}$$

This can be interpretted as the problem of how to decide the number of bots to combine with V_g. The right hand side is the cost of the additional bot. The left contains two terms. The first is the benefit of what the bot will pay: $A(\mu)$. But this benefit is reduced by the amount of lost attention from making the message get reliable, $A'(\mu)$, as μ falls with more bot traffic; the term in parenthesis is the amount of benefit generated per unit of good traffic, and as bot traffic changes by one unit, good traffic changes proportionally by a factor of μ^2. Simply put, the left hand side is the benefit of one additional bot, written in terms of μ. At an interior solution, as μ falls (i.e. bot traffic rises), this goes from being greater than the cost γ to less. Mechanically, the lower is γ, the more bot traffic can be absorbed before the left hand side balances the right. Intuitively, the cheaper is bot traffic, the more of it is worthwhile to run.

In particular, Proposition 5 and its proof highlight that there are parameters values such that $\mu^*(\overline{\theta}) = 1$ for a given value of γ but $\mu^*(\overline{\theta}) < 1$ for some $\gamma' < \gamma$. So falling ad revenue can result in platforms abandoning perfect certification, as was perhaps the case for Twitter.

Costs of targeted views. We conduct a similar comparative static for the cost c of directing interested views at good types. Again, there is a natural interpretation—more information on viewers, improved algorithms and analytics have likely reduced costs of targetting interested viewers. To consider the effect of such changes, we introduce a parameter $\kappa > 0$ (that only appears in the following discussion) such that the cost is $\kappa c(V_g)$.

Proposition 6. *Let the cost $\kappa c(V_g)$ of interested views V_g be parametrized by $\kappa > 0$. For all $\theta \in \Theta$, an increase in κ implies the following for all $\theta \in \Theta$:*

(i) *The quality $\mu^*(\theta)$ does not change.*
(ii) *The quantity of views $V_g^*(\theta)$ weakly decreases.*
(iii) *The set of values $\{\theta \in \Theta \mid V_g^*(\theta) > 0\}$ that are served does not change.*

Proposition 6 argues that when targetting improves (that is κ falls) so that it becomes cheaper to find engaged consumers for good content providers, the quality of certificates does not change; instead good content providers enjoy more viewers and the platform enjoys more bot revenue in proportion. From the first order condition for V_g, it is immediate that, since μ is constant as κ changes, views increase in θ at a rate proportional to $1\backslash\kappa$. As targetting costs decline, content becomes more skewed to high θ providers.

Shape of consumer attention function. Lastly, we also examine the effect of making the attention function more concave or convex. Intuitively, the concavity of consumer attention as a function of the quality of a certificate governs how a platform chooses certificate quality to trade off earnings from selling engaged views to good providers and what it can earn from bot revenues. For a fixed type of good provider, the only way to sell to more bots is to offer a worse certificate but this entails lower engaged views for the good provider (and even from the bots' perspective, reducing quality through sales to bots reduces attention). Starting from perfect certification, this cost of polluting good engagement is less pronounced for concave than convex attention functions and as a result, leads to lower quality certification.

This kind of concavity or convexity of attention reflects consumer preferences (in our model captured by the costs of reading posts, but intuitively in a more flexible formulation one might imagine this also reflecting anticipated benefits from good posts or harms from bad posts). Concavity is consistent with consumers who are particulalry keen to find good content and suffer relatively little from the inconvenience associated with looking at some bot content. Instead, those with convex attention can be understood as being harmed by even a little bad content. Different kinds of media content might be thougt of as differing on this scale, where scrolling past bad entertainment content is perhaps only an incovnenience whereas consuming fake news is more harmful. To the extent that this captures features of these different kinds of social media, the following result suggests a greater extent of bot traffic on entertainment-oriented social media than news-related media.[19]

Proposition 7. *Suppose that $\hat{A}(\mu) = g(A(\mu))$ for some increasing, differentiable, concave (convex) g with $g(0) = 0$ and $g(1) = 1$. Then the optimal μ is weakly lower (resp. higher) under $\hat{A}(\mu)$ than under $A(\mu)$.*

To get intuition on how a concave tranformation changes μ, consider the concave transformation $g(A) = min\{\alpha A, 1\}$ for $\alpha > 1$. Certainly if μ is past the point where $g(A(\mu)) = 1$, then μ is lower under the concave tranformation, since there is no reason for the platform to ever provide

[19] Of course, this is a *ceteris paribus* statement and one might expect, for example, that bots value views differently across these different kind of media, for example.

certification greater than than the lowest μ which provides $A(\mu) = 1$. But consider μ where $g(A(\mu)) < 1$. We see in (4) that the tranformation scales both terms by α, and therefore is the same as lowering costs by a factor of $1/\alpha$, and it has been shown that lower γ leads to lower μ. Essentially the benefits of bot traffic are scaled upward by a constant fraction.

More generally, for concave g, the first term in (4) is scaled by the average transformation $g(A)/A$, since it is the measure of total engagement, while the second is scaled by the marginal transformation $g'(A)$, since it is the marginal change in engagement. But for concave functions from the unit interval to the unit interval, the average is greater than the marginal. The intution is similar to that of γ: the concave transformation scales the benefits of bot traffic more than proportionally to its cost, and so the platform sells to more bots.

4.3. *Comparison to perfect certification*

Since regulators have suggested that consumers would benefit from enforced perfect certification, it is useful to compare the optimal contract with the one studied in the perfect certification benchmark. We can make the comparison ϕ-by-ϕ; since we have made no assumption on F (other than increasing virtual valuations), there is no way to draw an overall conclusion in general. Nonetheless, we can give a clear picture of the tradeoff of enforced perfect certification, and later show examples in γ and $A()$ where the resolution is unambiguous for any F. Perfect certification has a trade off across different value of ϕ, if there is any effect at all:

Proposition 8. *Recall that under perfect certification, $V_g(\theta) > 0$ if and only if $\phi(\theta) > \gamma$. With certification for sale, either:*

(1) $\mu^*(\theta) = 1$ *for all θ such that* $\phi(\theta) \geq \gamma$ *and* $V_g^*(\theta) = 0$ *for all θ with* $\phi(\theta) \leq \gamma$
(2) $\mu^*(\theta) < 1$ *for some θ with* $\phi(\theta) > \gamma$ *and* $V_g^*(\theta) > 0$ *for some θ with* $\phi(\theta) < \gamma$

Proof. Let $\pi(\phi)$ denote the contribution to profits generated by type with virtual valuation ϕ under profit maximization, and $\pi^{pc}(\phi)$ the profits under perfect certification. Both are continuous in ϕ by the theorem of the maximum. They are also increasing since any choice for ϕ, if mimicked for $\phi' > \phi$, generates higher profits at ϕ' and strictly if $V_g(\phi) > 0$. Since $\pi(\phi) \geq \pi^{pc}(\phi)$, it must be the case that $V_g^* > 0$ whenever $V_g^{pc} > 0$ There are therefore two possibilities in addition to cases 1 and 2 in the statement. There is also:

1(b) $\mu^*(\phi) = 1$ for all $\phi \geq \gamma$ and $V_g^* > 0$ for some $\phi \leq \gamma$

2(b) $\mu^*(\phi) < 1$ for some $\phi > \gamma$ and $V_g^* = 0$ for all $\phi < \gamma$

If $\mu^*(\phi) = 1$ for all $\phi > \gamma$, then $\pi(\gamma) = \pi^{pc}(\gamma) = 0$, i.e. $V_g^*(\phi) = 0$ for $\phi < \gamma$. In other words, case 1(b) is impossible.

For cases 2 and 2(b), suppose that $\mu^*(\phi) < 1$ for some $a > \phi > \gamma$. A sufficient condition for case 2 is that $\pi(\gamma) > 0$, since if this holds, then by continuity profits are strictly positive for some $\phi < \gamma$, and therefore $V_g > 0$ for some $\phi < \gamma$. For $a > \phi > \gamma$, since $\mu(\phi) = 1$ is not optimal, it must be that $\pi(\phi) > 0$. Therefore case 2b requires that $\pi(\phi) > 0$ for some $\phi > \gamma$ but $lim_{\phi\downarrow\gamma}\pi(\phi) = 0$. But then, since profits are continuous in ϕ, $\pi(\gamma) = 0$. But then $V_g(\gamma) = 0$ must be optimal, and therefore $\mu^*(\gamma) = 1$. But since μ^* is increasing, this implies in fact it is case 1 not case 2(b).

□

Since, under perfect certification, views are positive for $\phi > \gamma$, Proposition 8 implies that enforced perfect certification either doesn't matter (in case 1) or has a trade-off (in case 2): it might make consumers better off by reducing bots for ϕ that remain served under perfect certification, but always at a cost to the set of types served.

The optimal contract always has to serve at least as many types as perfect certification since the optimal contract can pick $\mu = 1$. That imperfect certification would lead to more ϕ being served for $A'(1)$ small enough is immediate from the one certificate result. This result goes further by showing that, if perfect certification matters, *strictly* more types are served when imperfect certification is allowed, regardless of $A()$. The result also implies that there will always be distributions F for which one contract may be better for consumers on average, given the rest of the parameters of the environment, since the bulk of the probability could be focused on either range. Below we depict this in an example with linear attention.

Note that, in the region that remains served under perfect certification, the benefits are not unambiguous: there still may be lower V_g under perfect certification. Indeed since profits are positive in case 2 for $\gamma = \phi$, this implies that $A(\mu^*)V_g^*$ is positive and therefore, since both are increasing, for $\phi > \gamma$ but sufficiently close, engagement is higher than under perfect certification, since views and therefore engagement are nearly zero near γ under perfect certification.

This trade off is particularly stark if the cost of targeting is zero up to some $\bar{V}_g$ and infinity after. In that case, any time $V_g(\theta) > 0$, it must be that $V_g(\theta) = \bar{V}_g$, since scaling V_g and V_b by any factor simply increases profits by that factor, so all types that are served are served maximally under either contract, and therefore the only effects are that consumers prefer better certification (fewer bots) and more types served. This shuts off the ambiguity for $\phi > \gamma$ in case 2: perfect certification has the same views and higher certification. This implies that, in case 2, regulating certification comes at a cost (types served below $\phi^{pc}_{min} = \gamma$ are not under perfect certification) as well as a benefit (with good type views fixed at $\bar{V}_g$, imperfect certification above ϕ^{pc}_{min} is worse for consumers than perfect certification). Depending on whether F puts more weight on values below or just above γ, the aggregate welfare effect will go one way or the other.

Small γ. To show how these forces might resolve, and because there is concern that many platforms have access to bad content very cheaply, consider the limit as the cost of untargeted views γ goes to zero, so that the platform can flood viewers with bots at low cost. Assume that targeting is still necessary for good content.[20] We focus on the case where the function $A(\mu)$ that governs how attention depends on the quality of the certificate is a power function; that is, $A(\mu) = \mu^{\alpha}$. We call a platform with $\alpha < 1$ concave, and $\alpha > 1$ convex. As suggested above, concavity leads to higher quality certificates and a diminished incentive to use good providers as a tool for generating bot revenue. But in this example that difference leads to extreme differences in the platform's structure as γ goes to zero.

In the limiting case where γ goes to zero, we show that the distinction between concavity and convexity is substantive in the following sense. We find that in this small γ case concave platforms always perform worse, in terms of engagement, than perfect certification would, but that convex platforms may perform better. In other words, regulation of a concave platform in the face of cheap bot traffic that enforces perfect certification would improve engagement, but the same regulation on a convex platform might be counterproductive.

We first show that the concave platform only generates engagement, in the limit, for virtual values higher than one—that is, some good content providers with virtual valuations for engagement that is higher than zero still receive no engagement, as they are completely flooded with bots, because the bots value being seen more than the content providers value engagement. In any case where engagement is positive, it converges to perfect certification. [21]

Proposition 9. *Suppose $A(\mu) = \mu^{\alpha}$ for $\alpha \leq 1$. Under perfect certification, $lim_{\gamma\to 0} V_g(\theta) > 0$ if and only if $\phi(\theta) > 0$. Then the solution to (2) has*

$$lim_{\gamma\to 0} A(\mu^*(\theta))V^*_g(\theta) = \begin{cases} 0 & \phi(\theta) < \bar{\phi} \\ c'^{-1}(\phi(\theta)) & \phi(\theta) > \bar{\phi} \end{cases}$$

where $\bar{\phi} \geq 1$.

[20] Throughout we assume that the probability of finding an interested user with an untargeted view is small relative to the costs of sending out an untargeted view. As we take the cost of untargeted views γ to zero, we therefore maintain that the probability that an untargeted view reaches an interested viewer, λ, goes to zero at the same rate as γ. That is $\frac{\gamma}{\lambda}$ is not falling, and targeting is still necessary for good providers.

[21] Note that the fact that any platform with concave A has zero engagement for $\phi\theta < 1$ is a consequence of taking the limit in the linear A case described above, and the fact that μ decreases under concave transformation of A

As γ goes to zero, types below $\phi = 1$ have such a low fraction of good types that they generate negligible engagement even if they are served, because of the temptation of the platform to sell views to bad content. For comparison, the solution under perfect certification has $lim_{\gamma \to 0} V_g = c'^{-1}(max\{0, \phi\})$; This implies that, for $\alpha < 1$ and γ small enough, there is more engagement under perfect certification for all but a vanishing set of types; types $\phi < \bar{\phi}$ are flooded with bots when bot traffic is cheap. Although case 2 of Proposition 8 applies, for all ϕ between zero and $\bar{\phi}$, perfect certification would improve engagement immensely, and for $\phi < \gamma$, the gains from content for sale are becoming negligible as γ goes to zero because certification is very low. In other words, in terms of engagement, perfect certification eventually dominates allowing certification for sale for concave $A()$ and small enough γ.

The same conclusion about engagement going to zero does not apply for convex platforms. Suppose $A(\mu) = \mu^\alpha$ for $\alpha > 1$. It is direct from the FOC that $A(\mu) = 1$ for $\phi > 1/\alpha$, but for $\phi < 1/\alpha$, for any ϕ where profits are feasibly positive, it must be that case that either $\mu = 1$ or the interior solution $\mu = \frac{1-1/\alpha}{1-\phi}$ applies, including engagement for $\phi < 0$, unlike under perfect certification. In other words, case 2 of Proposition 8 applies and the benefits of additional types served does not vanish.[22] Even as γ gets small, a convex platform may outperform enforced perfect certification, while a concave platform cannot; there is once again a trade off. We conclude that enforced perfect certification is a beneficial policy for a concave platform with cheap bot traffic for sale, but not necessarily for a convex platform under the same circumstance.

Linear Attention. Consumer welfare is proportional to engagement with linear attention, and therefore can be computed explicitly. For high enough θ, certification is perfect when for sale, and so welfare is equivalent. For a range of the lowest θ, where $\phi(\theta) < \gamma$, content is shown that is not shown under perfect certification, a welfare gain. For an intermediate range, however, the imperfect certification has offsetting effects; for θ close to perfect certification, the lost engagement from $\mu < 1$ more than offsets the higher V_g under the optimal contract and welfare is lower under content moderation for sale then under enforced perfect certification.

Since the bulk of the mass of types could be focused on any of the regions of the picture, welfare could be higher or lower with certification for sale. However, if for instance the support of F is entirely on the region where welfare gains are positive, clearly certification for sale is better than enforced perfect certification. As a general statement, if $\overline{\theta} < \gamma$, then enforced perfect certification leads to no good traffic on the platform, but certification for sale can still lead to good content being shown, and therefore perfect certification is not valuable for consumers. Enough low value ϕ providers make perfect certification perform worse; very high valuations make the two the

[22] When α grows large, the convex platform follows the same rule as perfect certification for $\phi > 0$; i.e. it converges to case 1 of the Proposition.

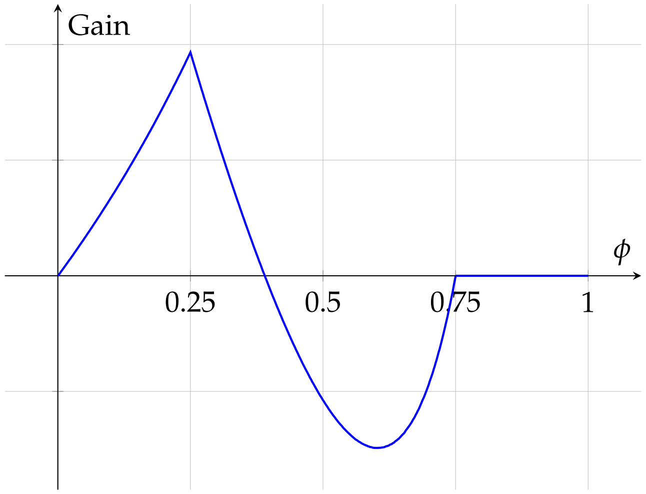


FIGURE 3. Welfare Gains from Certification for Sale: $A(\mu) = \mu, \gamma = 1/4, c(x) = x^2/2$

same, since the platform also enforces perfectly. The benefits of perfect certification come for intermediate values of ϕ.

5. Extension: Addiction and Losses from Bad Content

In this section, we extend the model to consider greater harms from bots and addiction. In the base model, the consumer engages in content if $\mu - q > 0$. In this section we consider adapting this consumer engagement problem in two ways in order to assess two policy-related ideas. First that bad content may be getting worse for consumers, and second that consumers may suffer from digital addiction. We separately implement both concerns into the model in rather straightforward ways and find intuitive results—as bot content becomes worse, platforms implement cleaner certification; and as consumers are more addict certification becomes worse and consumers are exposed to more bot traffic.

First, suppose that bad content generates losses b rather than a payoff of zero. In this case, a consumer engages in content when

$$\mu - (1-\mu)b - q > 0$$

i.e.

$$q < \mu - (1-\mu)b.$$

In effect this transforms the attention function to become $\hat{A}(\mu) = A(\mu(1+b) - b)$.

Second, we consider the possibility that consumers are addicted and lose a if they don't read. Then they read if $\mu - q > -a$, i.e. $q < \mu + a$. So the attention fucntion for those with addiction $a > 0$ becomes $\hat{A}(\mu) = A(\mu + a)$

We can assess these two possibilities through the following result.

Lemma 2. *Let* $\hat{A}(\mu) = A(h(\mu))$ *with* $\frac{A'(\mu)h'(\mu)}{A(h(\mu))} < (>)1$*. Then the optimal* μ *is lower (resp. higher) under* $\hat{A}(\mu)$ *than under* $A(\mu)$*.*

The lemma corroborates the claims made at the start of this section and immediately establishes the following results about the optimal μ as bad content becomes costly and users are addicted.

Corollary 2. $\mu^*(\theta)$ is increasing in b and decreasing in a

That is, platform content is cleaner when costs of bad content are higher and less clean as consumers are exposed to more bot content when they are addicted.

Finally, note that since this analysis follows the same approach to $A()$ as the rest of the paper, other results are unchanged. In particular Proposition 8 remains: even if $b > 0$, either enforced certification does not matter, or it increases the set of types served. Increasing the set of types served has to be good for consumers, even with $b > 0$, since they chose to read the posts even though they faced the costs b; the expected payoff was positive. Increasing b might change the platform's choice to perfect certification, but it does not change the conclusion that enforced perfect certification is either irrelevant, or comes with a trade off.[23]

6. Conclusion

Our analysis above builds on several themes which we developed through the benchmark cases before seeing them come to the fore in the characterization of the optimal mechanism.

First, and perhaps most familiar that platforms a platform can benefit from imperfect certification since doing so enables the platform to earn revenues from bad content but it must be mindful that diluting user experience in this way comes at a cost—in our model, this comes through reduced consumer attention.

Second, we illustrate how combining offers that include both different numbers of views and differing levels of certification can be more profitable for a platform—it optimally "pollutes" the certification of lower-value good quality providers from whom it would, in any case, be able to extract relatively little while maintaining perfect certification of higher-valuation providers and

[23] When consumers are addicted, revealed preference arguments are more tenuous.

therefore not sacrificing any revenues from them. Moreover, using different degrees of imperfect certification coupled with different levels of exposure allows a platform to better price discriminate.

Third, we show that consumers might benefit from imperfect certification relative to perfect certification through two channels. First, some good providers with low valuation who would not appear on the platform under perfect certification do in fact garner views under imperfect certification—in essence, the platform subsidizes their presence in order to earn from low-quality providers. Second, under imperfect certification, among those good content providers who are featured viewership is more egalitarian and less sensitive to the content provider's valuation than is the case under perfect certification. In this way, it brings it viewership closer to the solution to the engagement-maximizing solution which treats all good content identically. As the linear example illustrates, there are cases where consumers are strictly better off from the optimal mechanism than limiting platforms to perfect certification—in contrast to the tenor of some policy discussion.

The model builds off a familiar Mussa and Rosen (1978) framework, and its tractability (particularly when parameterized) allows us to examine some natural comparative statics and develop further results. Specifically, we highlighted that a lower ad revenue (a natural interpretation of the opportunity cost of allocating viewership) leads to a dilution of the quality of certificates, or, equivalently, more traffic being assigned to bots than to good content. Improved targetting raises platform profits and views assigned but has no impact on the quality of certificates that good content providers receive. Convexity of consumers' attention plays an important role where more convex attention leads to "purer" certificates; this convexity can be understood as capturing the extent to which consumers are willing to put up with bad content to enjoy good content with more convexity suggesting less patience with bad quality. Its role is brought into sharp relief in the example where the cost of untargeted views becomes vanishingly small.

Our results speak to an ongoing discussion surrounding content moderation in online platforms. Most obviously, our findings suggest that, in principle at least, consumers can benefit from allowing some bad content since it can be used to subsidize more good content and lead to more egalitarian content provision. It is also noteworthy, that the platform to some extent internalizes the harm associated with bad quality content—it makes consumers pay less attention, and so from the platform's perspective limits its ability to raise revenue and so the platform will issue purer certification if harms are higher. At an extreme, if harm is sufficiently high then trivially no consumers will engage and perfect certification will arise with no need for regulatory intervention. There may be alternative reasons for regulatory intervention—most obviously, consumer protection for naive consumers, and the possibility of externalities as discussed, for example in Bursztyn, Handel, Jimenez, and Roth (2023)—though a thorough examination lies beyond the scope of this paper.

Although we focus on the application to an online platform moderating content of good and bad type providers, there are other applications that fit the structure we introduce. One interpretation is that the bad type content is merely any hidden advertisement the platform can introduce. One can imagine a search platform that can put hidden ads among the organic search results, and separate them from the explicit advertisements. [24] Our model constructs the optimal way to mix these hidden ads into content, and highlights the potential costs of enforcing a lack of mixing of content.

A final interpretation is that the hidden advertisements are chosen by the good type content provider, but regulated by an outside force like the platform. An influencer can decide how much content to show that matches their own tastes, and therefore what their followers seek, and how much is not in their followers interest but is paid. In that case, the certificate can stand in for a form of disclosure regulation: perfect disclosure regulation corresponds to announcing the type of content post by post, and may not be optimal when steering is for sale. Imperfect disclosure, as often seems to arise, can be better than perfectly enforced disclosure regulations.

[24] Although Goggle does not explicitly charge for organic traffic, some have argued that having ad business with Google influences organic placement. Similarly Amazon favors suppliers that also purchase ancillary services.

Appendix A. Omitted proofs from the text

Proof of Lemma 1

Lemma 3. *Take any incentive compatible and individually rational mechanism (V_g, V_b, M, P) with associated quality μ. There exists another incentive compatible and individually rational mechanism $(\tilde{V}_g, \tilde{V}_b, \tilde{M}, P)$ such that, for all $\theta \in \Theta$, $\tilde{M}(\theta) = \theta$ and, both the platform and good providers receive the same payoff as from (V_g, V_b, M, P).*

Proof. Given a mechanism (V_g, V_b, M, P), we will construct another mechanism $(V_g, \tilde{V}_b, \tilde{M}, P)$ with the desired properties stated above.

First, we define $\tilde{M}(\theta) = \theta$. If $\mu(M(\theta)) = 0$, we define

$$\tilde{V}_b(\theta) = V_b(\theta) \text{ and } \tilde{V}_g(\theta) = 0 \qquad \text{for all } \theta \in \Theta.$$

Conversely, if $\mu(M(\theta)) > 0$, we define

$$\tilde{V}_b(\theta) = \frac{1-\mu(M(\theta))}{\mu(M(\theta))} V_g(\theta) \quad \text{and} \quad \tilde{V}_g(\theta) = V_g(\theta) \qquad \text{for all } \theta \in \Theta. \tag{5}$$

Let $\tilde{\mu}$ be the quality associated with mechanism $(\tilde{V}_g, \tilde{V}_b, \tilde{M}, P)$. Observe that, by construction,

$$\tilde{\mu}(\tilde{M}(\theta)) = \mu(M(\theta)) \qquad \text{for all } \theta \in \Theta. \tag{6}$$

Now note that

$$\theta A(\tilde{\mu}(\tilde{M}(\theta')))\tilde{V}_g(\theta') - P(\theta') = \theta A(\mu(M(\theta')))V_g(\theta') - P(\theta') \qquad \text{for all } \theta, \theta' \in \Theta.$$

In words, good providers of all values have the same payoff as the original mechanism (whether they report truthfully or misreport) and consequently $(\tilde{V}_g, \tilde{V}_b, \tilde{M}, P)$ is incentive compatible and individually rational because the original mechanism (V_g, V_b, M, P) is both.

Take an m in the image of M. If $\mu(m) = 0$, then

$$\mathbb{E}[V_b(\theta)|M(\theta) = m] = \mathbb{E}[\tilde{V}_b(\theta)|M(\theta) = m]$$

since V_b and $\tilde{V}_b$ are defined to be equal for such θ. When $\mu(m) > 0$, (5) and (6) together imply that

$$\begin{aligned}\mathbb{E}[V_b(\theta)|M(\theta) = m] = \frac{1-\mu(m)}{\mu(m)}\mathbb{E}[V_g(\theta)|M(\theta) = m] &= \mathbb{E}\left[\frac{1-\tilde{\mu}(\tilde{M}(\theta))}{\tilde{\mu}(\tilde{M}(\theta))}V_g(\theta)\middle| M(\theta) = m\right] \\ &= \mathbb{E}[\tilde{V}_b(\theta)|M(\theta) = m].\end{aligned}$$

Consequently,

$$\int_{\Theta} [A(\mu(M(\theta)))V_b(\theta) - \gamma V_b(\theta)] f(\theta)d\theta = \int_{\Theta} [A(\tilde{\mu}(\tilde{M}(\theta)))\tilde{V}_b(\theta) - \gamma \tilde{V}_b(\theta)] f(\theta)d\theta$$

and so the platform makes the identical profit from (V_g, V_b, M, P) and $(\tilde{V}_g, \tilde{V}_b, \tilde{M}, P)$ as required. □

Proposition 1. *There exists $\delta > 0$ such that, if $A'(1) < \delta$, there exists μ with $\phi_l(\mu) < \phi_l(1)$*

Proof. For $\phi_l(1) = \gamma + c'(0)$, consider decreasing μ slightly from 1. This type gets views when μ decreases if the costs on the left hand side of the first order condition increase more slowly than the benefits on the right, i.e. if

$$\gamma < 1 - (\gamma + c'(0))A'(1)$$

or

$$\gamma < \frac{1 - A'(1)c'(0)}{1 + A'(1)}$$

Since $\gamma < 1$, if $A'(1)$ is small enough, this holds. □

Proposition 2. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then the optimal single certification argmax$_{\bar{\mu}}\Pi^s(\bar{\mu}) <$ 1*

Proof. Since $\overline{\theta} \geq 1$, it must be the case that providing perfect certification is profitable for the platform since the virtual value of good types exceeds the marginal cost of providing views when views are small. The derivative of the payoff is

$$\frac{\partial \Pi^s(\bar{\mu})}{\partial \bar{\mu}} = \int_0^{\overline{\theta}} \left[\left(\phi(\theta) + \frac{1-\bar{\mu}}{\bar{\mu}} \right) A'(\bar{\mu}) - \frac{A(\bar{\mu})}{\bar{\mu}^2} + \frac{\gamma}{\bar{\mu}^2} \right] V_g(\theta) f(\theta) d\theta.$$

which evaluated at $\bar{\mu} = 1$ is $\int_0^{\overline{\theta}} [\phi(\theta)A'(1) - 1 + \gamma] V_g(\theta)f(\theta)d\theta$. Since $\phi(\overline{\theta})A'(1) < 1 - \gamma$, $A'(1) > 0$ and $\phi(\theta)$ is increasing, it is immediate that the derivative is negative at $\bar{\mu} = 1$. □

Proposition 3. *Suppose $\overline{\theta} \geq 1$, $\phi(\overline{\theta})A'(1) < 1 - \gamma$, and $\phi(\theta)$ is strictly increasing. Then the optimal solution to max$_{\underline{\mu},\overline{\mu},\hat{\theta}^{bin}}\Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin})$ has $\underline{\mu} < \overline{\mu}$.*

Proof. Let $h(\theta) = (\phi(\theta) + \frac{1-\bar{\mu}}{\bar{\mu}})A'(\bar{\mu}) - \frac{A(\bar{\mu})}{\bar{\mu}^2} + \frac{\gamma}{\bar{\mu}^2}$. Suppose $\underline{\mu} = \overline{\mu}$ is optimal. Then by Proposition 1, $0 < \underline{\mu} = \overline{\mu} < 1$. Now since h is strictly increasing, it must be the case that for a fixed μ either $h(\theta) < 0$ for all $\theta < \hat{\theta}^{bin}$ or $h(\theta) > 0$ for all $\theta > \hat{\theta}^{bin}$. In the former case, $\underline{\mu}$ cannot be optimal, since the derivative of $\Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin})$ is $\int_0^{\hat{\theta}^{bin}} h(\theta)f(\theta)d\theta < 0$ and the objective could be raised by decreasing $\underline{\mu}$. In the latter case, the reverse applies and $\overline{\mu}$ should be raised. □

Proposition 4. *There is an optimal mechanism (V_g^*, μ^*) solving the platform's problem (2) where both V_g^*, μ^* are nondecreasing and satisfy*

$$\mu^*(\theta) = \max\left\{\tilde{\mu} \;\middle|\; \tilde{\mu} \in \operatorname*{argmax}_{\hat{\mu}\in[0,1]}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\}\right\} \;>\; 0,$$
$$V_g^*(\theta) = c'^{-1}\left(\max\left\{\left[\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right]A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)}, 0\right\}\right)$$

for all $\theta \in \Theta$.

Proof.

We maximize the objective function pointwise and show that the mechanism we obtain satisfies the necessary monotonicity properties to satisfy the incentive compatibility constraints.

First, observe that, if

$$\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}} \geq \left(\phi(\theta) + \frac{1-\hat{\mu}'}{\hat{\mu}'}\right)A(\hat{\mu}') - \frac{\gamma}{\hat{\mu}'} \tag{7}$$

then, for any $V_g(\theta) \in \mathbb{R}_+$, the value of the objective function satisfies

$$\left[\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right]V_g(\theta) - c(V_g(\theta)) \geq \left[\left(\phi(\theta) + \frac{1-\hat{\mu}'}{\hat{\mu}'}\right)A(\hat{\mu}') - \frac{\gamma}{\hat{\mu}'}\right]V_g(\theta) - c(V_g(\theta))$$

and vice versa when inequality (7) is reversed.

Consequently, there is a pointwise optimum that satisfies

$$\mu^*(\theta) \in \operatorname*{argmax}_{\hat{\mu}\in[0,1]}\left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\}.$$

We pick $\mu^*(\theta)$ to be the largest above maximizer. Note that

$$\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}} = (\phi(\theta)-1)A(\hat{\mu}) + \frac{A(\hat{\mu})-\gamma}{\hat{\mu}} \quad\longrightarrow\quad -\infty$$

as $\hat{\mu} \to 0$ (because $A(\hat{\mu}) \to 0$) and therefore $\mu^*(\theta) > 0$ for all $\theta \in \Theta$.

We now argue that μ^* is nondecreasing. By definition,

$$\left(\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right)A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)} \geq \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}$$

for all $0 < \hat{\mu} < \mu^*(\theta)$. Then for all $\theta' > \theta$, it must be the case that

$$\left(\phi(\theta') + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right)A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)} \geq \left(\phi(\theta') + \frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}$$

for all $0 < \hat{\mu} < \mu^*(\theta)$ since $A(\mu^*(\theta)) > A(\hat{\mu})$ and ϕ is nondecreasing. Consequently, we must have $\mu^*(\theta') \geq \mu^*(\theta)$.

The function V_g^* as defined in the statement of the proposition is the solution to

$$V_g^*(\theta) = \operatorname*{argmax}_{v_g \in \mathbb{R}_+} \left\{ \left[\left(\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)} \right) A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)} \right] v_g - c(v_g) \right\} \tag{8}$$

and the exact expression is obtained from the first-order condition.

Now observe that, because ϕ is nondecreasing, the function

$$\left(\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)} \right) A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)} = \max_{\hat{\mu} \in [0,1]} \left\{ \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}} \right\}$$

is nondecreasing because it is the maximum of nondecreasing functions. This, in turn, implies from (8), that V^* is nondecreasing. Therefore $A(\mu^*(\cdot))V_g^*(\cdot)$ is nondecreasing and consequently the pointwise solution is incentive compatible as required. This completes the proof. □

Proposition 5. *For all $\theta \in \Theta$, the quality $\mu^*(\theta)$ is nondecreasing in γ.*

Proof. For clarity, in this proof, we use the notation $\mu_\gamma^*(\theta)$ to make the dependence on γ explicit.

Recall that $\mu_\gamma^*(\theta) > 0$ for all $\theta \in \Theta$. By the definition of μ_γ^* from Proposition 4, we have

$$\begin{aligned} &\left(\phi(\theta) + \frac{1-\mu_\gamma^*(\theta)}{\mu_\gamma^*(\theta)} \right) A(\mu_\gamma^*(\theta)) - \frac{\gamma}{\mu_\gamma^*(\theta)} \geq \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}} \\ \iff \gamma &\left(\frac{1}{\hat{\mu}} - \frac{1}{\mu_\gamma^*(\theta)} \right) \geq \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \left(\phi(\theta) + \frac{1-\mu_\gamma^*(\theta)}{\mu_\gamma^*(\theta)} \right) A(\mu_\gamma^*(\theta)) \end{aligned}$$

for all $\hat{\mu} \in (0, \mu_\gamma^*(\theta))$.

Therefore, for any $\gamma' > \gamma$, we must have

$$\begin{aligned} \gamma' &\left(\frac{1}{\hat{\mu}} - \frac{1}{\mu_\gamma^*(\theta)} \right) > \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \left(\phi(\theta) + \frac{1-\mu_\gamma^*(\theta)}{\mu_\gamma^*(\theta)} \right) A(\mu_\gamma^*(\theta)) \\ \iff &\left(\phi(\theta) + \frac{1-\mu_\gamma^*(\theta)}{\mu_\gamma^*(\theta)} \right) A(\mu_\gamma^*(\theta)) - \frac{\gamma'}{\mu_\gamma^*(\theta)} > \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) - \frac{\gamma'}{\hat{\mu}} \end{aligned}$$

for all $\hat{\mu} \in (0, \mu_\gamma^*(\theta))$. Consequently, we must have $\mu_{\gamma'}^*(\theta) \geq \mu_\gamma^*(\theta)$ as required. □

Proposition 6. *Let the cost $\kappa c(V_g)$ of interested views V_g be parametrized by $\kappa > 0$. For all $\theta \in \Theta$, an increase in κ implies the following for all $\theta \in \Theta$:*

(i) *The quality $\mu^*(\theta)$ does not change.*

(ii) *The set of values $\{\theta \in \Theta \mid V_g^*(\theta) > 0\}$ that are served does not change.*

(iii) *The quantity of views $V_g^*(\theta)$ weakly decreases.*

Proof. The first statement follows immediately from Proposition 4 since the cost $\kappa c(\cdot)$ does not enter the expression for μ^*.

From Proposition 4, $V_g^*(\theta)$ solves

$$V_g^*(\theta) = c^{-1}\left(\frac{1}{\kappa}\max\left\{\left[\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right]A(\mu^*(\theta)) - \frac{\gamma}{\mu^*(\theta)}, 0\right\}\right).$$

The second and third statements follow immediately from this equation. Note that $V_g^*(\theta) > 0$ whenever the right hand side of the above equation is greater than 0 and the sign of right hand side does not depend on κ. Clearly the term in the round brackets is nonincreasing in κ so $V_g^*(\theta)$ must be weakly decreasing. □

Proposition 7. *Suppose that $\hat{A}(\mu) = g(A(\mu))$ for some increasing, differentiable, concave (convex) g with $g(0) = 0$ and $g(1) = 1$. Then the optimal μ is weakly lower (resp. higher) under $\hat{A}(\mu)$ than under $A(\mu)$.*

Proof. Note that concavity (convexity) is equivalent to $\frac{Ag'(A)}{g(A)} < (>)1$ for any A since $g(0) = 0$. Moreover, concavity (convexity), $g(0) = 0$, and $g(1) = 1$ implies $\hat{A}(\mu) > (<)A(\mu)$

The FOC for μ is

$$\left(\phi - \frac{1-\mu}{\mu}\right)\hat{A}'(\mu) - \frac{\hat{A}(\mu)}{\mu^2} + \frac{\gamma}{\mu^2} = 0$$
$$\left(\phi - \frac{1-\mu}{\mu}\right)\frac{\hat{A}'(\mu)}{\hat{A}(\mu)} - \frac{1}{\mu^2} + \frac{\gamma}{\hat{A}(\mu)\mu^2} = 0$$
$$\left(\phi - \frac{1-\mu}{\mu}\right)\frac{g'A'(\mu)}{g(A)} - \frac{1}{\mu^2} + \frac{\gamma}{\hat{A}(\mu)\mu^2} = 0$$

Suppose g is concave. For any μ, the last term is smaller than under $\hat{A}$ since $\hat{A} > A$ and the first term is smaller if $\frac{g'A'(\mu)}{g(A)} < \frac{A'(\mu)}{A}$,which is true if $\frac{Ag'(A)}{g(A)} < 1$. Therefore μ is smaller. The reverse holds for g convex. □

Proposition 8. *Suppose $A(\mu) = \mu^\alpha$ for $\alpha \leq 1$. Then the solution to (2) has*

$$lim_{\gamma\to 0}A(\mu)V_g = \begin{cases} 0 & \phi < \bar{\phi} \\ c'^{-1}(\phi) & \phi > \bar{\phi} \end{cases}$$

where $\bar{\phi} \geq 1$.

Proof. Consider some ϕ where the solution has $A(\mu) > 0$ for some finite $\bar{\gamma} > 0$. (Otherwise profits are zero for this ϕ and without loss $A(\mu)V_g = 0$.) This implies that the solution has $\mu > 0$ for any finite $\gamma < \bar{\gamma}$ since profits are positive and therefore μ cannot be zero. The FOC is

$$(\phi\mu + 1 - \mu)\mu A'(\mu) - A(\mu) + \gamma = 0$$

so

$$\mu^{\alpha}\left((\phi\mu + 1 - \mu)\alpha - 1\right) + \gamma = 0$$

Note that either the term in brackets is zero or μ is zero for γ small; i.e. the interior solution is

$$\mu = \frac{1/\alpha - 1}{\phi - 1}.$$

But next we show that the SOC cannot be satisfied for $\phi > 1$. The SOC requires that

$$\begin{aligned} \alpha\mu^{\alpha-1}\left((\phi\mu + 1 - \mu)\alpha - 1\right) + \alpha\mu^{\alpha}(\phi - 1) &< 0 \\ \frac{\alpha}{\mu}\left(\mu^{\alpha}\left((\phi\mu + 1 - \mu)\alpha - 1\right) + (\phi - 1)\right) &< 0 \\ \frac{\alpha}{\mu}\left(-\gamma + (\phi - 1)\right) &< 0 \end{aligned}$$

Where the last line replaces with the first term with the FOC. For γ small, this cannot hold for $\phi > 1$.

The conclusion is that the interior solution applies only if $\phi < 1$, but the interior solution is only positive for $\phi < 1$ if $\alpha > 1$. As a result, for γ small, for $\alpha < 1$ all μ are converging to either zero or one. □

Lemma 4. *Let $\hat{A}(\mu) = A(h(\mu))$ with $\frac{A'(\mu)h'(\mu)}{A(h(\mu))} < (>)1$. Then the optimal μ is lower (resp. higher) under $\hat{A}(\mu)$ than under $A(\mu)$.*

Proof. Then FOC for μ

$$\begin{aligned} \left(\phi - \frac{1-\mu}{\mu}\right)\frac{\hat{A}'(\mu)}{\hat{A}(\mu)} - \frac{\gamma}{\mu^2} &= 0 \\ \left(\phi - \frac{1-\mu}{\mu}\right)\frac{A'(\mu)h'(\mu)}{A(h(\mu))} - \frac{\gamma}{\mu^2} &= 0 \end{aligned}$$

Notice that μ is lower if $\frac{A'(\mu)h'(\mu)}{A(h(\mu))} < \frac{A'(\mu)}{A(\mu)}$,which is true if $\frac{A(\mu)h'(\mu)}{A(h(\mu))} < 1$ the result holds. □

Appendix B. Two certificates analysis

Here we provide a characterization of the platform's problem when restricted to two certificates.

Proposition 8. *Consider the version of the platform's problem (2) where $\mu : \Theta \to \{\underline{\mu}, \overline{\mu}\}$ and $V_g : \Theta \to \mathbb{R}_+$.*

There is an optimal mechanism (V_b^{bin}, μ^{bin}) *given by*

$$\hat{\theta} = \begin{cases} \min\left\{\theta \in \Theta \;\middle|\; \left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}}\right) A(\overline{\mu}) - \frac{\gamma}{\overline{\mu}} \geq \left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}}\right) A(\underline{\mu}) - \frac{\gamma}{\underline{\mu}}\right\} & \text{if the set is non-empty,} \\ \overline{\theta} & \text{otherwise,} \end{cases}$$

$$\mu^{bin}(\theta) = \begin{cases} \overline{\mu} & \text{if } \theta > \hat{\theta} \text{ or } \theta = \hat{\theta} < \overline{\theta}, \\ \underline{\mu} & \text{if } \theta < \hat{\theta} \text{ or } \theta = \hat{\theta} = \overline{\theta}, \end{cases} \qquad \text{and}$$

$$V_g^{bin}(\theta) = c'^{-1}\left(\max\left\{\left[\phi(\theta) + \frac{1-\mu(\theta)}{\mu(\theta)}\right] A(\mu(\theta)) - \frac{\gamma}{\mu(\theta)}, 0\right\}\right).$$

Proof. We maximize the objective function pointwise and show that the mechanism we obtain satisfies the necessary monotonicity properties to satisfy the incentive compatibility constraints.

First, observe that, if

$$\left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}}\right) A(\overline{\mu}) - \frac{\gamma}{\overline{\mu}} \geq \left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}}\right) A(\underline{\mu}) - \frac{\gamma}{\underline{\mu}} \tag{9}$$

then, for any $V_g(\theta) \in \mathbb{R}_+$, the value of the objective function satisfies

$$\left[\left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}}\right) A(\overline{\mu}) - \frac{\gamma}{\overline{\mu}}\right] V_g(\theta) - c(V_g(\theta)) \geq \left[\left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}}\right) A(\underline{\mu}) - \frac{\gamma}{\underline{\mu}}\right] V_g(\theta) - c(V_g(\theta))$$

and vice versa when inequality (9) is reversed.

Consequently, there is a pointwise optimum that satisfies

$$\mu^{bin}(\theta) \in \underset{\hat{\mu} \in \{\underline{\mu}, \overline{\mu}\}}{\operatorname{argmax}} \left\{\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}\right\}.$$

For any $\theta \in \Theta$ at which both $\underline{\mu}$ and $\overline{\mu}$ are maximizers, we set $\mu^{bin}(\theta) = \overline{\mu}$.

Now note that, if (9) holds for some θ, it also holds for all $\theta' > \theta$. This is because $A(\overline{\mu}) > A(\underline{\mu})$ and $\phi(\cdot)$ is nondecreasing. Consequently, μ^{bin} must take the cutoff form

$$\mu^{bin}(\theta) = \begin{cases} \overline{\mu} & \text{when } \theta > \hat{\theta}, \\ \underline{\mu} & \text{when } \theta < \hat{\theta}. \end{cases}$$

as in the statement of the proposition and thus μ^{bin} is nondecreasing.

The function V_g^{bin} as defined in the statement of the proposition is the solution to

$$V_g^{bin}(\theta) = \underset{v_g \in \mathbb{R}_+}{\operatorname{argmax}} \left\{\left[\left(\phi(\theta) + \frac{1-\mu^{bin}(\theta)}{\mu^{bin}(\theta)}\right) A(\mu^{bin}(\theta)) - \frac{\gamma}{\mu^{bin}(\theta)}\right] v_g - c(v_g)\right\} \tag{10}$$

and the exact expression is obtained from the first-order condition.

Now observe that, because ϕ is nondecreasing, the function

$$\left(\phi(\theta)+\frac{1-\mu^{bin}(\theta)}{\mu^{bin}(\theta)}\right)A(\mu^{bin}(\theta))-\frac{\gamma}{\mu^{bin}(\theta)}=\max_{\hat{\mu}\in\{\underline{\mu},\overline{\mu}\}}\left\{\left(\phi(\theta)+\frac{1-\hat{\mu}}{\hat{\mu}}\right)A(\hat{\mu})-\frac{\gamma}{\hat{\mu}}\right\}$$

is nondecreasing because it is the maximum of two nondecreasing functions. This, in turn, implies from (10), that V^{bin} is nondecreasing. Therefore $A(\mu^{bin}(\cdot))V_g^{bin}(\cdot)$ is nondecreasing and consequently the pointwise solution is incentive compatible as required. This completes the proof. □

References

Acemoglu, D., et al. (2023): "Content moderation and public discourse," *Journal of Political Economy*, 131(4), 1120–1148.

Ali, S. N., N. Haghpanah, X. Lin, and R. Siegel (2022): "How to sell hard information," *The Quarterly Journal of Economics*, 137(1), 619–678.

Aridor, G., R. Jiménez-Durán, R. Levy, and L. Song (forthcoming): "The Economics of Social Media," *Journal of Economic Literature*.

Armstrong, M., and J. Zhou (2011): "Paying for prominence," *The Economic Journal*, 121(556), F368–F395.

Athey, S., and E. Ellison (2011): "Position auctions with consumer search," *The Quarterly Journal of Economics*, 126(3), 1213–1270.

Bar-Isaac, H., and S. Shelegia (2022): *Monetizing steering*. Centre for Economic Policy Research.

Böhme, E. (2016): "Second-degree price discrimination on two-sided markets," *Review of Network Economics*, 15(2), 91–115.

Bouvard, M., and R. Levy (2018): "Two-sided reputation in certification markets," *Management Science*, 64(10), 4755–4774.

Burguet, R., R. Caminal, and M. Ellman (2015): "In Google we trust?," *International Journal of Industrial Organization*, 39, 44–55.

Bursztyn, L., B. R. Handel, R. Jimenez, and C. Roth (2023): "When product markets become collective traps: The case of social media," Discussion paper, National Bureau of Economic Research.

Chen, Y., and C. He (2011): "Paid placement: Advertising and search on the internet," *The Economic Journal*, 121(556), F309–F328.

Choi, J. P., D.-S. Jeon, and B.-C. Kim (2015): "Net neutrality, business models, and internet interconnection," *American Economic Journal: Microeconomics*, 7(3), 104–141.

Competition and Markets Authority (2022): "Auditing algorithms: the existing landscape, role of regulators and future outlook," accessed on June 7, 2024 at https://www.gov.uk/government/publications/findings-from-the-drcf-algorithmic-processing-workstream-spring-2022/auditing-algorithms-the-existing-landscape-role-of-regulators-and-future-outlook.

Corts, K. S. (2013): "Prohibitions on false and unsubstantiated claims: Inducing the acquisition and revelation of information through competition policy," *The Journal of Law and Economics*, 56(2), 453–486.

——— (2014): "Finite optimal penalties for false advertising," *The Journal of Industrial Economics*, 62(4), 661–681.

DE CORNIERE, A., AND G. TAYLOR (2014): "Integration and search engine bias," *The RAND Journal of Economics*, 45(3), 576–597.

——— (2019): "A model of biased intermediation," *The RAND Journal of Economics*, 50(4), 854–882.

DENECKERE, R. J., AND R. PRESTON MCAFEE (1996): "Damaged goods," *Journal of Economics & Management Strategy*, 5(2), 149–174.

DRANOVE, D., AND G. Z. JIN (2010): "Quality disclosure and certification: Theory and practice," *Journal of economic literature*, 48(4), 935–963.

EDELMAN, B., M. OSTROVSKY, AND M. SCHWARZ (2007): "Internet advertising and the generalized second-price auction: Selling billions of dollars worth of keywords," *American economic review*, 97(1), 242–259.

ERSHOV, D., AND M. MITCHELL (Forthcoming): "The Effects of Advertising Disclosure Regulations on Social Media: Evidence From Instagram," *RAND Journal of Economics*, Forthcoming.

FAINMESSER, I. P., AND A. GALEOTTI (2021): "The Market for Online Influence," *American Economic Journal: Microeconomics*, 13(3), 1–28.

GLAESER, E. L., AND G. UJHELYI (2010): "Regulating misinformation," *Journal of public Economics*, 94(3-4), 247–257.

GOLDFARB, A., AND C. TUCKER (2019): "Digital economics," *Journal of Economic Literature*, 57(1), 3–43.

HAGIU, A., AND B. JULLIEN (2014): "Search diversion and platform competition," *International Journal of Industrial Organization*, 33, 48–60.

ICHIHASHI, S., AND A. SMOLIN (2023): "Buyer-Optimal Algorithmic Consumption," SSRN Working Paper, September 21, 2023.

INDERST, R., AND M. OTTAVIANI (2012): "Competition through Commissions and Kickbacks," *American Economic Review*, 102(2), 780–809.

JEON, D.-S., B.-C. KIM, AND D. MENICUCCI (2022): "Second-degree price discrimination by a two-sided monopoly platform," *American Economic Journal: Microeconomics*, 14(2), 322–369.

JEON, D.-S., AND S. LOVO (2013): "Credit rating industry: A helicopter tour of stylized facts and recent theories," *International Journal of Industrial Organization*, 31(5), 643–651.

JOHNSON, J. P., A. RHODES, AND M. WILDENBEEST (2023): "Platform Design When Sellers Use Pricing Algorithms," *Econometrica*, 91(5), 1841–1879.

KOMINERS, S. D., AND J. M. SHAPIRO (2024): "Content Moderation with Opaque Policies," Working Paper w32156, National Bureau of Economic Research (NBER), Available at SSRN: https://ssrn.com/abstract=4731065.

LIZZERI, A. (1999): "Information revelation and certification intermediaries," *The RAND Journal of Economics*, pp. 214–231.

MADIO, L., AND M. QUINN (2024): "Content Moderation and Advertising in Social Media Platforms," Working Paper 11169, CESifo Working Paper, Available at SSRN: https://ssrn.com/abstract=4875546 or http://dx.doi.org/10.2139/ssrn.4875546.

MITCHELL, M. (2021): "Free ad (vice): internet influencers and disclosure regulation," *The RAND Journal of Economics*, 52(1), 3–21.

MUSSA, M., AND S. ROSEN (1978): "Monopoly and product quality," *Journal of Economic theory*, 18(2), 301–317.

RHODES, D., AND J. WILSON (2018): "False Advertising," *RAND Journal of Economics*, 49(4), 1011–1039.

SRINIVASAN, K. (2023): "Paying Attention," Discussion paper, Mimeo.

WHITE, L. J. (2010): "Markets: The credit rating agencies," *Journal of Economic Perspectives*, 24(2), 211–226.

ZOU, T., Y. WU, AND M. SARVARY (2025): "Designing Recommendation Systems on Content Platforms: Trading off Quality and Variety," *Available at SSRN*.