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: April 18, 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. This informs recent policy discussion around the allowing of selling of certification on platforms.

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 it price discriminates 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

[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.

is an economic force acting against this intuition that can overwhelm it. Specifically, consumers 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 bad content is 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 content 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 low quality traffic; 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 results 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 traffic 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 trade-off 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. In the next section, we review the literature. Then, we introduce the model and construct and simplify 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 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 affect attention in the platform's profit maximization 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 profit-maximizing for the certifier 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 (advertiser) 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 content providers of fixed type).

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*ood or *b*ad; good content can be of value to interested consumers (described below), but bad content cannot. There is a continuum of a unit mass of good providers whose private *values* $\theta \in [0, \bar{\theta}] =: \Theta$ are distributed according to $F \in \Delta(\Theta)$ 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 good 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 $\theta a v_g$.

Consumers never engage with content from bad providers and thus they only value their content being read. A bad content provider receives utility av_b from $v_b \in \mathbb{R}_+$ views that pay a level $a \in [0, 1]$ of attention to their content.

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. Alternatively, the assumption is consistent with the idea that, since

[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]Given our assumptions on the costs of the platform that follow, we could equivalently assume that there is a fixed mass of bad types but that their content can be spread widely.

[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 bad providers since the platform can sell to those with highest value.

bad type content is never of interest to the consumer, a single piece could be reused and shown to many consumers.

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 direct advertisements at consumers instead of content.

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. There are a unit mass of consumers who may view content. Each piece of content is accompanied by a message m.[11] Consumers observe the message m and decide whether to read the content to learn more it. Reading requires paying a time cost $q \geq 0$, distributed according to a strictly increasing, differentiable cumulative distribution function $A(q)$. The consumer receives a payoff of 1 if they decide to read the content and it turns out to be good content that matches their interest (such content generates engagement). They receive a payoff of 0 otherwise.

Upon observing a message m, consumers assign probability $\hat{\mu} \in [0, 1]$ to the content being good and of interest. The expected payoff from reading the content is $\hat{\mu} - q$. Consequently, a consumer chooses to read the content if, and only if, $q \leq \hat{\mu}$. In other words, $A(\hat{\mu})$ is the likelihood that content labeled with message m is read by consumers. We henceforth refer to A as the *attention function*. Since engagement is the consumer's payoff, the engagement maximizing benchmark is a natural consumer welfare standard that we analyze below. Aside from allowing a welfare calculation, the

[8]For example, platforms can distinguish between good and bad providers exactly as users can but need only do so once on behalf of all potential viewers. Further, in addition to observing content directly, platforms monitor consumer reactions and other measures of engagement.

[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 with the hope of randomly finding interested users. Suppose that the unconditional probability of a user being interested in an untargeted post is λ. The optimal mechanism we derive using the cost function c will remain optimal for a sufficiently low λ. This is because it is cheaper to target than to pay the opportunity cost of the number of untargeted posts required to generate 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.

details of the underlying interpretation of the attention function as a cost calculation embodied in q is inessential. The same predictions arise, under any functional relationship between beliefs and attention that satisfies our assumptions (that is, a strictly increasing, differentiable, cummulative distribution function).

Platform Pricing. The platform price discriminates by offering a (direct) *mechanism* to providers. The mechanism consists of four functions. These correspond to the message or *certificate* assigned to a good provider claiming to be of type θ; the number of targeted views that this provider receives; the number of untargeted bad provider views that are assigned to type θ's certificate; and the price that the good provider pays to receive the certificate. As discussed below, the surplus from bad providers is fully extracted and so we do not need incorporate their payments into the mechanism.

These functions are written as follows:

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

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. Additionally, there are $V_b(\theta)$ untargeted views by bad providers that are also assigned certificated $M(\theta)$.

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 sums 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 $A(\mu(M(\theta)))V_b(\theta)$. $A(\mu(M(\theta)))$ is the fraction of consumers who pay attention to content marked with a certificate $M(\theta)$. The cost of $V_b(\theta)$ untargeted views is $\gamma V_b(\theta)$ and the cost of $V_g(\theta)$ targeted views is $c(V_g(\theta)) + \gamma V_g(\theta)$ and has an additional component associated with targetting. The constraint captures both incentive compatibility and individual rationality constraints for good providers.

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 than the value of the views. Therefore, we henceforth focus on the more interesting case $A(1) > \gamma$ and we normalize $A(1) = 1$.

, and so henceforth we focus on the more interesting case in which $A(1) > \gamma$.

2.1. *Preliminary Analysis and Simplification 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 receiving it. 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 $(V_g, \tilde{V}_b, \tilde{M}, P)$ 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).*

In words, this lemma 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. The intuition is related to the linearity of the payoff in

V_b; the planner can reallocate V_b freely across any pooled message to make the ratio of good to bad content equal to the average level, type-by-type.[12]

This allows us to drop the certification decision M and the function V_b determining the bad provider views and rewrite the platform's problem with it choosing the quality $\mu : \Theta \to [0,1]$ of the certificate for each good provider value θ. This is a consequence of Lemma 1 and the fact that the bad provider views V_b are pinned down by the equation $\mu(\theta) = V_g(\theta)/[V_g(\theta) + V_b(\theta)]$. The platform thus solves[13]

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

subject to

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

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

$$\max_{V_g,\mu} \int_\Theta \left[\left(\phi(\theta) + \frac{1-\mu(\theta)}{\mu(\theta)} \right) A(\mu(\theta))V_g(\theta) - c(V_g(\theta)) - \gamma\frac{V_g(\theta)}{\mu(\theta)} \right] f(\theta)d\theta, \tag{2}$$

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 observing that it is optimal for the platform to provide a utility of zero to a good provider of value $\theta = 0$. We refer to a solution of the platform's problem (2) as an *optimal mechanism*.

The following lemma summarizes the simplification of the planner's problem. It states that a pointwise solution that satisfies the monotonicity required for incentive compatibility is the optimal mechanism.

[12] If the payoff to bad-type views were not linear, there might be incentive to pool messages across θ to smooth V_b.

[13] In what follows, note that, in the integrand, a fraction whose numerator and denominator are both zero takes the value zero. We also note that, as written, the problem permits us to choose $\mu(\theta) > 0$ and $V_g(\theta) = 0$ (which technically violates the definition of μ). 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.

Lemma 2. *Suppose there is a mechanism (V_g^p, μ^p) such that*

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

pointwise maximizes the platform's objective for all $\theta \in \Theta$ and that $A(\mu^p(\cdot))V_g^p(\cdot)$ is nondecreasing. Then (V_g^p, μ^p) is an optimal mechanism; that is, it is a solution to the platform's problem (2).

In what follows, we assume the distribution F is such that the virtual value ϕ is strictly increasing. This is a standard technical assumption in mechanism design. As we will show, this assumption simplifies the derivation of the optimal mechanism; it can be obtained by pointwise maximization as described in Lemma 2. This assumption can be dispensed with by employing standard ironing techniques but we choose not to do so because we view the additional technicality (required for this generality) complicates the presentation of our main economic insights.

3. BENCHMARKS

In this section, we derive three benchmarks that provide context for the properties of the optimal mechanism. We first solve a "planner's problem" in which the goal is to maximize consumer engagement net of platform costs. We then derive the properties of the optimal mechanism when the platform is restricted to using a constant quality for all Θ (which includes not mixing good and bad types, as desired by the European Commission, when quality is perfect). This contrasts the maximization of consumer engagement and platform profits. We then derive the properties of the optimal mechanism when the platform is restricted to offering only two certificates (so the quality function can take at most two values). This allows us to transparently demonstrate the value to the platform of selling multiple certificates, which is typically a feature of the solution to the full problem described in the last section.

3.1. *The planner's problem*

Consider a planner who wants to maximize the level of engagement on the platform net of platform costs.[14] That is, they want to solve

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

[14]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.

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 content provider's utility which incorporates the value of that engagement and writes as $\theta A(\mu(\theta))V_g(\theta)$) and the remaining terms are the costs associated with directing both targeted to good providers and and untargeted views to bad ones respectively (recalling that $V_b(\theta) = \frac{V_g(\theta)}{\mu}$). Recall that views to bad providers do not generate engagement.

The solution (V_g^e, μ^e) to the above planner's problem is immediate from pointwise optimization (since there are no constraints) and is given by

$$\mu^e(\theta) = 1,$$
$$V_g^e(\theta) = c'^{-1}(1-\gamma)$$

for $\theta \in \Theta$.[15]

We flag two intuitive properties of this solution. The first is that no views are directed to bad providers since directing content to them is costly and does not generate engagement. 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 traffic to providers with higher willingness to pay for those views.

3.2. *One certificate*

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 platform (2), the function μ is restricted to be a constant function that takes some value $\hat{\mu} \in [0, 1]$.

We do this for several reasons. First, this corresponds to a realistic form of content moderation whereby platforms do not distinguish between different kinds of content (that is all content is presented in the same way) but still choose how many views to allocate to different providers. The platform may choose to ban bad provider traffic (perfect certification) or not (imperfect certification).[16] Second, the analysis crisply illustrates how imperfect certification can raise platform profits by expanding the subset of types Θ 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. Finally, it allows us to characterize the solution if the platform is forced to certify accurately; that is, to label good and bad types as such.

[15]Recall that we have normalized $A(1) = 1$.

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

We first derive the optimal choice of traffic V^g for an arbitrary quality $\hat{\mu} \in [0,1]$. We then vary the quality $\hat{\mu}$ 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).[17] 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

$$\begin{aligned} \Pi^s(\hat{\mu}) := \max_{V_g} & \left\{ \int \left(\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}) V_g(\theta) - \gamma \frac{V_g(\theta)}{\hat{\mu}} - c(V_g(\theta)) \right) f(\theta) d\theta \right\} \\ & \text{subject to} \\ & V_g(\theta) \geq V_g(\theta') \quad \text{for } \theta \geq \theta',\ \theta, \theta' \in \Theta. \end{aligned} \tag{3}$$

It is immediate here that the pointwise optimum

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

satisfies the required monotonicity constraint. For $\theta \in \Theta$ such that $V_g^s(\theta)$ is not zero, the value is derived from the first-order condition

$$\frac{\gamma}{\hat{\mu}} + c'(V_g^s(\theta)) = \left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}} \right) A(\hat{\mu}). \tag{4}$$

A useful special case 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.[18] Then, plugging in $\hat{\mu} = 1$, V_g^s is given by $V_g^s(\theta) = c'^{-1}(\max\{\phi(\theta) - \gamma, 0\})$.

This solution also provides a natural connection to classic price discrimination: it 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

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

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

valuation ϕ. Those good providers with higher θ who value being seen more reveal their higher valuation by paying more to enjoy a higher number of engaged views.

Following the first-order condition (4), imperfect certification has two effects. First, it lowers attention and therefore reduces the amount that can be charged to good types; this reduction is proportional to the virtual value. On the other hand, it generates, for every good view, some bad traffic that can be monetized. This effect is constant and raises the payoff from views symmetrically for all virtual values. Therefore imperfect certification increases the value of views relatively more for low virtual values. As a result, imperfect certification makes traffic more egalitarian.

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

Proposition 1. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then, the optimal single certification*

$$\operatorname*{argmax}_{\hat{\mu}} \{\Pi^s(\hat{\mu})\} < 1$$

is imperfect.

If $A'(1)$ is not too big, then the gains from imperfect certification outweigh the costs of lost attention and engagement. The return $1 - \gamma$ to bad provider traffic near perfect certification provides the sufficient bound on $A'(1)$.

Imperfect certification, when it is desirable for the platform, raises content diversity. Formally, we define the *content diversity* of a mechanism (V_g, μ) as the set of types $\{\theta \in \Theta | V_g(\theta) > 0\}$ that are allocated views. Thus, a mechanism (V_g', μ') has greater content diversity than another mechanism (V_g, μ) if a larger set of types are served by the former or $\{\theta \in \Theta | V_g'(\theta) > 0\} \supset \{\theta \in \Theta | V_g(\theta) > 0\}$. The following result shows that either the platform prefers perfect certification over a fixed level of imperfect certification, or the imperfect certification is serving to increase content diversity:

Proposition 2. *For any $0 < \hat{\mu} < 1$, either $\Pi^s(\hat{\mu}) < \Pi^s(1)$, or content diversity is higher under $\hat{\mu}$ then under perfect certification.*

This previews our result which will extend to the full mechanism below: either it is optimal for the platform to certify perfectly, or forcing perfect certification lowers content diversity.

Linear Attention. Throughout the paper, we will use the example $A(\mu) = \mu$ of a linear attention function (which corresponds to a uniform distribution of the cost of reading a post) to illustrate the underlying forces driving our results. In particular, here, we use it to demonstrate how greater content diversity can arise with imperfect certification.

For a fixed level of certification $\hat{\mu}$, the corresponding optimal V_g^s is given by

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

For perfect certification $\hat{\mu} = 1$ therefore, $V_g^s(\theta) = c^{-1}(\max\{\phi(\theta) - \gamma, 0\})$.

Figure 1 shows the number of views a good content provider enjoys (on the y-axis) as a function of their virtual value (and therefore, indirectly of their type) and compares perfect certification (depicted by the dashed red line) and imperfect certification with a single certificate of quality $\hat{\mu} = \frac{1}{2}$ (depicted by the blue line).

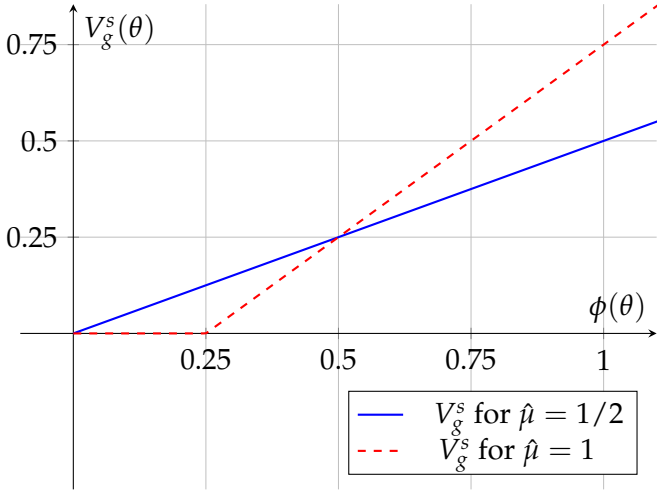


FIGURE 1. Good provider views with simple certification. Parameters: $A(\hat{\mu}) = \hat{\mu}$, $\gamma = 1/4$ and $c(v_g) = v_g^2/2$.

The figure highlights the two senses discussed above in which imperfect certification results in traffic that is closer to the engagement maximizing benchmark (recall that the engagement maximizing benchmark has identical traffic for all θ). First, note that there is greater content diversity under imperfect certification; specifically, more lower value good providers are served with imperfect certification. Second, traffic is more egalitarian in that the number of views are less sensitive to the virtual value (and therefore the type) with imperfect certification (observe that the red line corresponding to perfect certification is steeper). Of course, these notions of improved content diversity do not translate directly into higher welfare, which is captured by engagement. The welfare comparison is not immediately obvious from the figure since, in shifting from perfect to imperfect certification, some (higher) types receive fewer and other (lower) types receive more views. Finally, all types receive less attention conditional on being viewed. Morever, in this figure, we have not taken a stance on the distribution F of types. We take up welfare in more detail below. However it is immediate that if the mass of F is concentrated on lower values then welfare must be higher with imperfect certification.

3.3. *Two certificates*

Before moving on to our analysis of the general problem, we build some intuition by considering the case of two certificates. This case may also be of independent interest in that it mirrors recent developments in a couple of social media platforms. Specifically, 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.

The previous subsection showed that fixed imperfect (relative to perfect) certification can lead to higher profits and greater content diversity. The case with two certificates further illustrates how multiple levels of certification can raise the profits for the platform. By using two certificates, the platform can profitably serve low value good types, by cross subsidizing them with bad providers, while simultaneously not sacrificing engagement for the views of high value good types.[19] Furthermore, the two certificate benchmark allows us to demonstrate how 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 imperfect certification may be profitable, but is not fully optimal.

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 $\underline{\mu}, \overline{\mu} \in [0,1]$ with $\underline{\mu} \leq \overline{\mu}$. We write the two certificate benchmark problem for the platform as[20]

$$\begin{aligned}\Pi^{bin}(\underline{\mu}, \overline{\mu}, \hat{\theta}) := \max_{V_g} \Bigg\{ &\int_0^{\hat{\theta}} \left[\left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}} \right) A(\underline{\mu}) V_g(\theta) - c(V_g(\theta)) - \gamma \frac{V_g(\theta)}{\underline{\mu}} \right] f(\theta) d\theta \\ &+ \int_{\hat{\theta}}^{\overline{\theta}} \left[\left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}} \right) A(\overline{\mu}) V_g(\theta) - c(V_g(\theta)) - \gamma \frac{V_g(\theta)}{\overline{\mu}} \right] f(\theta) d\theta \Bigg\},\end{aligned} \tag{5}$$

[19] 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. This results in different economic effects.

[20] We characterize the two certificate optimal mechanism in Appendix B; we choose not present the result here for brevity. We show that the optimal two certificate mechanism takes the cutoff form assumed in the above problem. In other words, the optimal two certificate mechanism can be derived by choosing the appropriate $\underline{\mu}, \overline{\mu}$ and $\hat{\theta}$ to maximize $\Pi^{bin}(\underline{\mu}, \overline{\mu}, \hat{\theta})$.

where $\hat{\theta} \in \Theta$. In the above problem, all types above and below $\hat{\theta}$ are assigned the higher $\overline{\mu}$ and lower $\underline{\mu}$ quality certificates respectively.

In addition to higher value good types receiving the higher quality certificate, they also receive more targeted views. Thus, as in the single certificate optimum, the platform uses the quantity of traffic to price discriminate but now also uses the certificate quality to separate higher from lower value good types. The platform assigns higher-quality certificates to high-value good providers so as not to dilute earnings from these types, while assigning lower-quality certificates to use low-value good providers to use them as a means of earning revenue from the bad providers.

Instead of presenting the two certificate optimum, we derive a sufficient condition under which the platform gets strictly higher profits from using two certificates (as opposed to a single certificate $\underline{\mu} = \overline{\mu}$). In fact, this is the same sufficient condition under which the platform will choose imperfect certification when restricted to a single certificate (Proposition 1); namely, that attention does not drop too quickly as quality is reduced away from perfect certification.

Proposition 3. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then every two certificate optimum*

$$(\underline{\mu}^*, \overline{\mu}^*, \hat{\theta}^{bin^*}) \in \underset{\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin}}{\operatorname{argmax}} \left\{ \Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin}) \right\}$$

satisfies $\underline{\mu}^* < \overline{\mu}^*$.

When a single perfect certificate is not optimal (guaranteed by the condition in the above proposition), the platform will employ imperfect certification. The platform's problem (5) clearly demonstrates that the profit depends on the interaction of the buyer's type with the quality. Loosely speaking, since incentives are monotone in θ, targeting different quality levels to low and high types should improve profits. We next turn to the fully optimal mechanism in which any number of messages can be used. The forces outlined in both the single and two certificate cases come to the fore: the platform uses different quantities of views to price discriminate coupled with the assignment of distinct quality certificates to effectively extract revenue from bad providers.

4. Certification and steering for sale

4.1. *The optimal mechanism*

We now 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 bad providers and can more profitably price discriminate across good content providers by offering different quantities of targeted views

and distinct quality certificates. 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*

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

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

for all $\theta \in \Theta$.

Recall that incentive compatibility is satisfied if the product $A(\mu^*(\cdot))V_g^*(\cdot)$ is non-decreasing; thus, one 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 in Section 3.2 that pushed the optimal mechanism with imperfect (relative to perfect) certification qualitatively closer to planner optimal also apply in the optimal mechanism of Proposition 4. Serving low value good providers allows the platform to earn more revenue from bad providers. This, in turn, flattens out the relationship between targeted views and good provider valuations leading to more egalitarian content provision relative to perfect certification.

The optimal mechanism of Proposition 4 features greater content diversity than either the single or two certificate benchmarks. To see this, it is instructive to compare the optimal (unrestricted) mechanism to the mechanism (derived in Section 3.3) that maximizes profits when the platform is restricted to offer only two certificates. For 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. This greater content diversity is a result of the optimal mechanism 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 $A(\hat{\mu}) = \hat{\mu}$ is linear, Proposition 4 implies that the mechanism

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

$$V_g^*(\theta) = \begin{cases} c'^{-1}\left(\phi(\theta)\mu^*(\theta) + 1 - \mu^*(\theta) - \frac{\gamma}{\mu^*(\theta)}\right) & \text{if } \phi(\theta) \geq 1 - \frac{1}{4\gamma}, \\ 0 & \text{if } \phi(\theta) < 1 - \frac{1}{4\gamma} \end{cases}$$

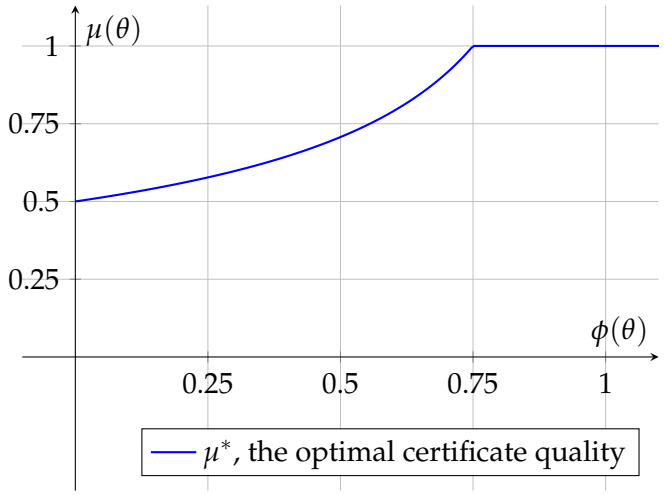
is optimal.

Now observe that $1 - \frac{1}{4\gamma} \geq 1 - \gamma$ when $\gamma \geq 1/2$. Consequently, when $\gamma \geq 1/2$, the optimal mechanism has the property that $\mu^*(\theta) = 1$ for all $\theta \in \Theta$ such that $V_g^*(\theta) > 0$. That is, perfect certification is optimal when it is sufficiently costly to supply untargeted views. We therefore focus on the case of $\gamma < 1/2$ in which the optimal mechanism features a variety of levels of certification.

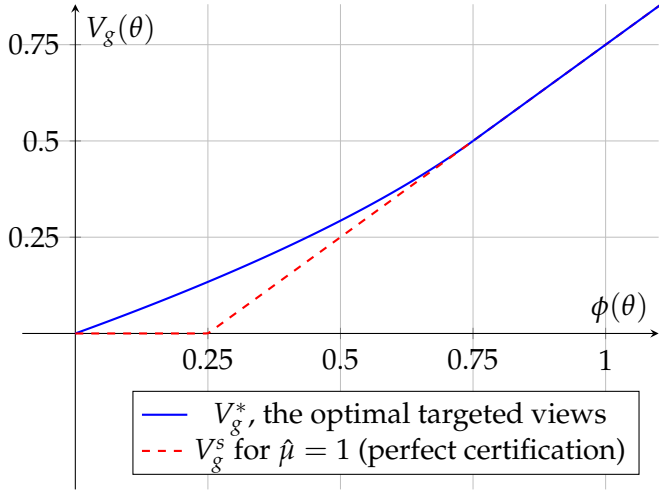
Figure 2 illustrates such a case. Each panel has the good provider's virtual value ϕ on the x-axis. Panel (A) plots the certificate quality μ^* and panel (B) plots the number of good provider views V_g^* in the optimal mechanism. Sufficiently high types are all assigned perfect certificates but the platform still price discriminates through the number of views it provides. The platform price discriminates using both instruments—imperfect certification and the quantity of views—for the lower types.

As a result of many certificates, content diversity increases relative to perfect certification, both in terms of types served, and the amount of views for good types of content 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, 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.



(A) Certificate quality



(B) Content Diversity

FIGURE 2. The optimal mechanism. Parameters: $A(\hat{\mu}) = \hat{\mu}, \gamma = 1/4, c(v_g) = v_g^2/2$.

4.2. *Comparative Statics*

The solution described in Proposition 4 allows us to conduct several comparative statics exercises. They help to explain the circumstances under which imperfect certification is optimal. Lower costs γ for untargeted view or a more concave attention function A make optimal certification more imperfect but a lower cost of targeting c has no impact.

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 next result shows that when ad revenue falls, good content providers enjoy worse certificates.

Proposition 5. *The quality $\mu^*(\theta)$ is nondecreasing in γ for all $\theta \in \Theta$.*

For intuition, consider a good provider with value $\theta \in \Theta$ that is assigned to an imperfect certificate or that $\mu^*(\theta) \in (0,1)$. Since, $\mu^*(\theta)$ maximizes $\left(\phi(\theta) + \frac{1-\hat{\mu}}{\hat{\mu}}\right) A(\hat{\mu}) - \frac{\gamma}{\hat{\mu}}$ (see Proposition 4), it satisfies the first-order condition

$$A(\mu^*(\theta)) - \mu^*(\theta)^2 \left(\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right) A'(\mu^*(\theta)) = \gamma. \tag{6}$$

This equation captures the tradeoffs of including additional bad content with a given certificate. The right hand side is the cost of the additional bad view. The left contains two terms. The first is the amount $A(\mu^*(\theta))$ that a bad type pays for this view. But including more bad content reduces the quality of the certificate and therefore the attention; this, in turn, reduces the amount that both good and bad providers are willing to pay. On the margin, additional bad views reduce the quality by $\frac{\mu^*(\theta)^2}{V_g^*(\theta)}$ and therefore the attention by $\frac{\mu^*(\theta)^2 A'(\mu^*(\theta))}{V_g^*(\theta)}$. The total amount paid for this certificate by both good and bad providers is $\left(\phi(\theta) + \frac{1-\mu^*(\theta)}{\mu^*(\theta)}\right) V_g^*(\theta)$. Therefore, on the margin, the lost revenue from a reduction of quality is given by the product of these two terms. If γ is lowered, more bad type traffic can be absorbed before the left hand side balances the right. Very intuitively, it is worthwhile to have more bad views if they are cheaper.

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 providers. Again, there is a natural interpretation: more information on viewers, improved algorithms and analytics reduce costs of targeting 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 of v_g targeted views to good providers 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 targeting 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 views and the platform enjoys more bad provider revenue in proportion. This comparative static is immediate from the expressions for the optimal mechanism in Proposition 4.

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 bad type revenues. For a fixed type of good provider, the only way to sell to more bad types is to offer a worse certificate but this entails lower engaged views for the good provider (and even from the bad types' perspective, reducing quality 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 content, but intuitively in a more flexible formulation one might imagine this also reflecting anticipated benefits from good content or harms from bad content). Concavity is consistent with consumers who are particularly keen to find good content and suffer relatively little from the inconvenience associated with looking at some bad 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 thought of as differing on this scale, where scrolling past bad entertainment content is perhaps only an inconvenience 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 bad traffic on entertainment-oriented social media than news-related media.[21]

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 certification greater than than the lowest μ which provides $A(\mu) = 1$. But consider μ where $g(A(\mu)) < 1$. We see in (6) 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 bad traffic are scaled upward by a constant fraction.

More generally, for concave g, the first term in Equation (6), corresponding to the first order condition for μ is scaled by the average transformation $g(A)/A$, since it is the measure of total

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

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 bad traffic more than proportionally to its cost, and so the platform sells to more bad types.

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 type-by-type (or equivalently 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 (notably small enough γ and concave $A(.)$) where the resolution is unambiguous for any F. If there is any effect at all, there is a tradeoff in the benefit of imperfect vs perfect certification across different values f ϕ as described in the following result.

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 $\phi(\theta) \leq \gamma$, i.e. the platform chooses perfect certification*
(2) *$\mu^*(\theta) < 1$ for some θ with $\phi(\theta) > \gamma$ and $V_g^*(\theta) > 0$ for some θ with $\phi(\theta) < \gamma$, i.e. enforced perfect certification reduces content diversity*

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 bad content 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 greater content diversity 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 potential benefits of perfect certification 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.[22]

Linear Attention. In order to see this tradeoff explicitly, and in order to connect to consumer welfare, we return to the linear case. 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.

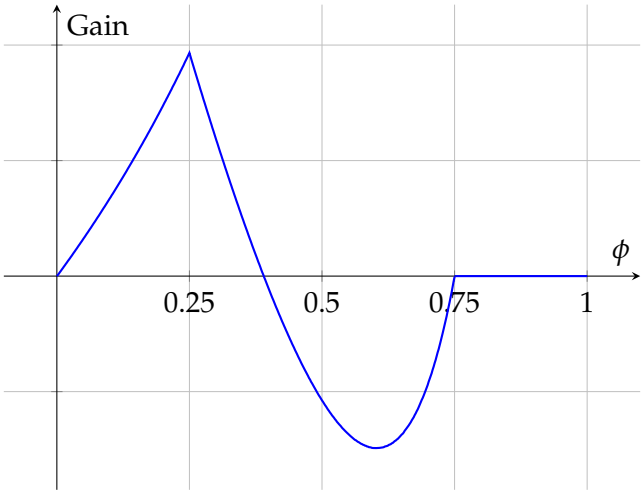


FIGURE 3. Welfare Gains from Certification for Sale: $A(\mu) = \mu, \gamma = 1/4, c(x) = x^2/2$

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

[22]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 bad types) and more types served. This shuts off the ambiguity for $\phi > \gamma$ in case 2: perfect certification has the same views and higher 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.

leads to no 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 θ providers make perfect certification perform worse; very high valuations make the two the same, since the platform also enforces perfectly. The benefits of perfect certification come for intermediate values of θ.

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 bad content at low cost. Assume that targeting is still necessary for good content.[23] 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. Proposition 8 implies that higher α leads to higher quality certificates and a diminished incentive to use good providers as a tool for generating bad type 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. When γ is small, 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 platformto enforce perfect certification in the face of cheap bad type traffic 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 as γ goes to zero, for virtual values higher than one, whereas prefect certification would generate engagement for all positive virtual values. In other words, some good content providers that would be served under perfect certification receive no engagement, as they are completely flooded with bad content. In any case where engagement is positive in the limit when certification is for sale, the allocation converges to perfect certification.[24]

[23]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.

[24]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

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 platform's profit maximizing strategy (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 $\bar{\phi} \geq 1$.

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 bad content. 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 $\alpha > 1$. Consider $\phi = 0$. It is direct from (6) that $\mu = 1 - 1/\alpha$, and therefore $V_g > 0$; i.e. there is positive engagement bounded away from zero even when γ is small. This in turn implies that engagement is bounded away from zero for small but negative ϕ, which would be unserved under perfect certification. In other words, case 2 of Proposition 8 applies and the benefits of additional types served does not vanish.[25] We conclude that enforced perfect certification is a beneficial policy for a concave platform with cheap bad traffic for sale, but not necessarily for a convex platform under the same circumstance.

5. Extension: Addiction and Losses from Bad Content

In this section, we extend the model to consider greater harms from bad types, as well as 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 bad content becomes worse, platforms

[25]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.

implement cleaner certification; and as consumers are more addicted certification becomes worse and consumers are exposed to more bad 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 $\hat{A}(\mu) = A(\mu + a)$

To understand the impact of a and b, consider the return to bad type views on the left hand side of the first order condition for μ from (6). For addiction, first suppose $\hat{A}(\mu) \in (0,1)$. An increase in a has no impact on the slope of the attention function and makes its level higher. This immediately implies that the level of μ increases. If $\hat{A}(\mu) = 1$, the platform chooses μ so that it is at the minumum level that attains $\hat{A}(\mu) = 1$. Since this value falls with a, μ must fall with a in all cases where attention is positive. On the other hand, for an increase in b, the return to bad content falls since the level of the attention function falls and its slope increases anywhere attention is positive. Therefore μ rises with b.

We summarize this with the following proposition.

Proposition 10. *Anywhere attention is positive, $\mu^*(\theta)$ is increasing in b and decreasing in a*

Platform content is cleaner when costs of bad content are higher and less clean as consumers are exposed to more bad 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 content 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.[26]

[26] Although, when consumers are addicted, revealed preference arguments are more tenuous.

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 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 bad content 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. [27] 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.

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

APPENDIX A. OMITTED PROOFS FROM THE TEXT

This section contains the proofs to all the results from the body of the paper. For ease of reference, we restate each results prior to presenting the proof.

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 $(V_g, \tilde{V}_b, \tilde{M}, P)$ 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

$$\tilde{V}_b(\theta) = \frac{1 - \mu(M(\theta))}{\mu(M(\theta))} V_g(\theta) \quad \text{and} \quad \tilde{V}_g(\theta) = V_g(\theta) \qquad \text{for all } \theta \in \Theta. \tag{7}$$

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

$$\tilde{\mu}(\tilde{M}(\theta)) = \mu(M(\theta)) \qquad \text{for all } \theta \in \Theta. \tag{8}$$

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 $\mu(m) > 0$, (7) and (8) 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,

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

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. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then, the optimal single certification*

$$\operatorname*{argmax}_{\hat{\mu}} \{\Pi^s(\hat{\mu})\} < 1$$

is imperfect.

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 $\int_0^{\overline{\theta}} \left[\phi(\theta)A'(1) - 1 + \gamma\right] V_g(\theta) f(\theta) d\theta$. 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 2. *For any $0 < \hat{\mu} < 1$, either $\Pi^s(\hat{\mu}) < \Pi^s(1)$, or content diversity is higher under $\hat{\mu}$ then under perfect certification.*

Proof. Define

$$R(\phi, \mu) = \left(\phi + \frac{1-\mu}{\mu}\right) A(\mu) - \gamma/\mu$$

The integrand in (3) can be written as $R(\phi(\theta), \mu)V_g - c(V_g)$. Therefore $V_g > 0$ if and only if $R(\phi, \mu) > 0$, i.e. for all $\phi > \underline{\phi}(\mu)$ where $R(\underline{\phi}(\mu), \mu) = 0$. Note that the derivative of $R(\phi, \mu)$ with respect to ϕ is $A(\mu) > 0$, so the function is linear and increasing in ϕ.

Next we compare $\mu = \hat{\mu} \in (0, 1)$ to $\mu = 1$. Since in both cases R is linear in ϕ with greater slope for $\mu = 1$, they cross exactly once, at $\bar{\phi}$. If $R(\bar{\phi}, \hat{\mu}) \leq 0$, then $R(\phi, 1) \geq R(\phi, \hat{\mu})$ for all ϕ with R positive, and therefore $\Pi^s(\hat{\mu}) < \Pi^s(1)$. On the other hand, if $R(\bar{\phi}, \hat{\mu}) > 0$, then $\underline{\phi}(\mu) = \bar{\phi} - \frac{R(\bar{\phi}, \mu)}{A(\mu)}$ and therefore $\underline{\phi}(\hat{\mu}) < \underline{\phi}(1)$, i.e. content diversity is higher under $\hat{\mu} < 1$ than when $\mu = 1$. □

Proposition 3. *Suppose $\overline{\theta} \geq 1$ and $\phi(\overline{\theta})A'(1) < 1 - \gamma$. Then every two certificate optimum*

$$(\underline{\mu}^*, \overline{\mu}^*, \hat{\theta}^{bin^*}) \in \underset{\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin}}{\operatorname{argmax}} \left\{ \Pi(\underline{\mu}, \overline{\mu}, \hat{\theta}^{bin}) \right\}$$

satisfies $\underline{\mu}^* < \overline{\mu}^*$.

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 2, $0 < \underline{\mu} = \overline{\mu} < 1$. 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*

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

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

for all $\theta \in \Theta$.

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

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

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 (9) is reversed.

Consequently, there is a pointwise optimum that satisfies

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

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

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

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

Now observe that, because ϕ is nondecreasing, the function

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

is nondecreasing because it is the maximum of nondecreasing functions. This, in turn, implies from (10), 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. *The quality $\mu^*(\theta)$ is nondecreasing in γ for all $\theta \in \Theta$.*

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

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

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

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 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.*

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. *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 $\phi(\theta) \leq \gamma$, i.e. the platform chooses perfect certification*
(2) *$\mu^*(\theta) < 1$ for some θ with $\phi(\theta) > \gamma$ and $V_g^*(\theta) > 0$ for some θ with $\phi(\theta) < \gamma$, i.e. enforced perfect certification reduces content diversity*

Proof. An exhaustive list of possibilities is case 1 and 2 plus:

(3) $\mu^*(\theta) = 1$ for all θ such that $\phi(\theta) \geq \gamma$ and $V_g^*(\theta) > 0$ for some θ such that $\phi(\theta) \leq \gamma$ (i.e. greater content diversity under certification for sale, and perfect certification whenever enforced perfect certification has views)

(4) $\mu^*(\theta) < 1$ for some θ such that $\phi(\theta) > \gamma$ and $V_g^*(\theta) = 0$ for all θ such that $\phi(\theta) < \gamma$ (i.e. no greater content diversity under certification for sale, and imperfect certification when enforced perfect certification has views)

We show that (3) and (4) cannot occur. Denote the contribution to profits by the type θ by

$$\pi(\theta) = max_V R(\phi(\theta), \mu^*(\theta))V - c(V)$$

Similarly denote the analogous contribution to profits under perfect certification by

$$\pi^{pc}(\theta) = max_V R(\phi(\theta), 1)V - c(V)$$

These are both continuous by the theorem of the maximum. They are increasing since any choice for θ, if mimicked for $\theta' > \theta$, generates higher profits at θ'.

Suppose Case 3. This requires that $\mu^*(\theta) = 1$ for all θ with $\phi(\theta) \geq \gamma$, so $\pi(\theta) = \pi^{pc}(\theta) = 0$ for $\phi(\theta) = \gamma$, and therefore $\pi(\theta) = V_g^*(\theta) = 0$ for $\phi(\theta) \leq \gamma$. This contradicts $V_g^*(\theta) > 0$ for some θ such that $\phi(\theta) \leq \gamma$ in Case 3.

Suppose Case 4. This requires $\pi(\theta) = 0$ for $\phi(\theta) < \gamma$, so $\pi(\theta) = 0$ for $\phi(\theta) = \gamma$ by continuity. Case 4 requires further that $\lim_{\phi(\theta)\downarrow\gamma} \mu^*(\theta) < 1$ and $\lim_{\phi(\theta)\downarrow\gamma} R(\phi(\theta), \mu^*(\theta)) = 0$, since if R (and therefore profits) were positive in the limit, profits would also be positive for θ below $\phi(\theta) = \gamma$. But $R(\theta, 1)$=0 when $\phi(\theta) = \gamma$, so $\mu^*(\theta) = 1$ for $\phi(\theta) = \gamma$. Since $\mu^*(\theta)$ is increasing, this contradicts $\lim_{\phi(\theta)\downarrow\gamma} \mu^*(\theta) < 1$.

□

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 platform's profit maximizing strategy (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 $\bar{\phi} \geq 1$.

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. □

APPENDIX B. TWO CERTIFICATES ANALYSIS

In this section, we provide a full characterization of the platform's optimal mechanism when restricted to two certificates.

Proposition 11. *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

$$\left(\phi(\theta) + \frac{1-\overline{\mu}}{\overline{\mu}}\right) A(\overline{\mu}) - \frac{\gamma}{\overline{\mu}} \geq \left(\phi(\theta) + \frac{1-\underline{\mu}}{\underline{\mu}}\right) A(\underline{\mu}) - \frac{\gamma}{\underline{\mu}} \tag{11}$$

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 (11) is reversed.

Consequently, there is a pointwise optimum that satisfies

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

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 (11) 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

$$V_g^{bin}(\theta) = \operatorname*{argmax}_{v_g \in \mathbb{R}_+} \left\{\left[\left(\phi(\theta) + \frac{1-\mu^{bin}(\theta)}{\mu^{bin}(\theta)}\right) A(\mu^{bin}(\theta)) - \frac{\gamma}{\mu^{bin}(\theta)}\right] v_g - c(v_g)\right\} \tag{12}$$

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

Now observe that, because ϕ is nondecreasing, the function

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

is nondecreasing because it is the maximum of two nondecreasing functions. This, in turn, implies from (12), 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*.