How to enforce platforms’ liability?

Adam Feher*

April 19, 2024

Abstract

New regulations are emerging to address the increase of illegal content on online platforms, highlighted by initiatives like the Digital Services Act. This paper presents a theoretical model examining how such regulations shift the economic incentives for social media platforms to moderate user-generated content. The model focuses on a strictly liable platform with heterogeneous users who may breach the rules set by a regulator. The larger the audience of the user, the higher the benefit she brings to the platform, but also the larger the societal harm if she violates the rules. The paper argues that oversimplified regulations may worsen the problem of cherry-picking, where platforms penalize only low type users and not high type ones for violations. To combat cherry-picking, one option for the regulator is to stop monitoring content that the platform has already removed, but this could lead to overmoderation. Another strategy involves the regulator enhancing its technology to monitor more content and adjusting fines based on the ex post observable size of the audience of users. Intriguingly, under this approach, the optimal fine might be reduced when more users commit violations.

Keywords— Content moderation, User-generated content, Platform, Digital Services Act

1 Introduction

The economic and societal relevance of online platforms has been on the rise. Platforms like Facebook, Twitter, and YouTube have attained immense popularity. According to COOK [2023], more than 5 billion YouTube videos are viewed daily, and Twitter publishes over 500 million tweets each day as of December 2023.¹ However, hosting platforms also contribute to disseminating illegal material, such as hate speech on social media platforms or copyright violations on media platforms. In response, today’s online intermediaries regulate users’ access

*University of Amsterdam and IAAEU, Email: feher@iaaeu.de
¹The daily number of tweets is reported by worldometers.info.
to platforms and police their behavior and expression on their platforms. Despite their efforts, the public pressure is mounting on platforms to do more to combat offenses committed under their supervision and on regulators to force platforms to police [Buiten et al., 2020]. The Directive on electronic commerce in the European Union and Section 230 in the United States Communications Decency Act granted immunity to online platforms with respect to third-party content for the past 20 years. However, these regulations are considered outdated and new laws and proposals have been made. The German Network Enforcement Act (Netzwerkdurchsetzungsgesetz, henceforth NetzDG) was enacted in 2018. Moreover, the Digital Services and Markets Act in Europe, the online safety bill in the UK, and a heated policy debate on reform and court cases in the US indicate an increased liability platforms are facing.

A fundamental question is how a regulator can enforce social media platforms’ liability for offenses committed by platform users. The key friction is that the platform has a technological advantage over the regulator in detecting offenses and may strategically use it to circumvent regulatory objectives. The theoretical model, which mirrors the German NetzDG, introduces a novel aspect wherein the platform engages in cherry-picking, to opportunistically punish select users on the platform. The platform’s aim is to create an illusion of compliance with regulatory standards and avoid sanctions. By selectively punishing certain users, the platform reduces the likelihood of the regulator detecting violations and subsequently penalizing the platform. Cherry-picking is particularly harmful if the platform provides preferential treatment to users with large audiences whose contributions to the platform’s profit are the largest, but whose offenses can cause the greatest societal damage.

Regulations, such as the NetzDG, that do not take cherry-picking into account can actually incentivize this behavior and decrease social welfare. To prevent cherry-picking, the regulator can implement a mechanism where the size of the fine depends on how the platform handles violations by users with small audiences compared to users with large audiences. The optimal schedule for sanctions can be nonmonotonic in the offenses the regulator detects on the platform. Specifically, the sanction may be higher when the regulator detects fewer violations on the platform. The reason is that from the regulator’s point of view, there are two punishable actions: the offense by the user and the cherry-picking by the platform.

A key example of the phenomenon depicted by the model is Twitch’s ’do-not-ban-list." This

---

2 In 2019, Facebook CEO Mark Zuckerberg declared that they would be allocating 5% of the firm revenues, 3.7 billion, on content moderation [Roettgers, 2019].

3 See, for example, the court cases and rulings of Bolger v. Amazon, Loomis v. Amazon, and Reed [2023] for legislative changes.

5 The idea that under certain liability regimes platforms may cherry-pick on what to remove (because risky) and what to maintain (because generating substantial revenues) first appears in Lefouilli and Madici [2022].
list reveals that top streamers were given special treatment when it came to suspending their streaming accounts [Grayson, 2021]. Another example is the Twitter posts from the account of President Donald Trump. Despite frequently sharing highly controversial content, his account remained active for an extended period, only being suspended following the tragic events at Capitol Hill [Andrews, 2020]. When Twitter did suspend the account, its share price plummeted by 12%, causing an approximate instantaneous loss of 5 billion dollars to the platform [Theron, 2021]. This decline in revenue can be attributed to a decrease in platform visits and subsequent advertising income, given that the account boasted one of the largest follower bases worldwide. This example highlights the misalignment of private incentives with social ones, as platforms may hesitate to take action against top users, potentially causing substantial societal harm.

An alternative narrative to the model is the extent to which (social) media platforms’ algorithms promote content that may be exploitative or harmful to certain groups. These algorithms prioritize content that generates higher engagement rates, such as views, comments, and likes, without necessarily considering the actual content of the message [Baluja et al., 2008]. As a result, the platform may promote content that titillates, shocks, enrages, divides, or is fake, aiming to exploit users’ cognitive limitations [Bhargava, 2022; Rosenquist et al., 2022; Montag et al., 2019; Bhargava and Velasquez, 2021]. Fake news is a notable example of such content [Van Alstyne et al., 2023]. This issue’s real-world implications are evident in recent legal proceedings. For instance, in Reynaldo Gonzalez v. Google, the plaintiff argues that Google’s content promotion on its platforms contributed to the radicalization of individuals, and subsequent terrorist attacks, making the company liable for the harm caused in the Paris 2015 attacks [McCabe, 2023]. Similarly, Meta is facing a lawsuit for allegedly disseminating personal information that ultimately led to an individual’s murder, despite receiving reports and requests for its removal [Perrigo, 2022]. In both cases, private incentives to promote harmful content may have outweighed the socially optimal level, resulting in significant societal harm.

The core of the theoretical model lies in the platform’s cherry-picking behavior. The platform maximizes its profit by selecting which user’s content to retain and which to delete if it does not adhere to the regulator’s rules. Users are heterogeneous; higher type users generate larger revenue for the platform, leading to a significant revenue loss if their content is deleted. When subjected to a flat fine and with the regulator monitoring all content having been published on the platform, the platform is incentivized to delete the content of low-type users to reduce the probability of the fine, while retaining top users’ content to boost revenue. Cherry-picking decreases social welfare if the platform removes content with positive social value, that is if

---

6 deleted or retained by the platform
the content of low types is socially valuable. Increasing the fine exacerbates the platform’s cherry-picking incentive, indicating that merely raising the fine does not solve the regulator’s problem, contrary to classical crime models like Becker [1968].

The regulator can employ alternative strategies in overseeing the content on the platform in order to improve social welfare. Specifically, the regulator can focus only on content that the platform chooses to retain. However, this approach also has its inherent drawbacks. For one, the platform might deliberately eliminate content with positive social value without any repercussions. The platform could intentionally remove socially valuable content without facing any consequences. For example, YouTube curtailed the visibility of videos pertaining to the Hong Kong protests in 2019 [Oremus, 2019]. In a separate incident, Twitter, in 2022, suspended several journalist accounts at the behest of Elon Musk [Sayantani, 2022]. Furthermore, any inadvertent removals by the platform would remain unchecked by the regulator, leading to potential issues. In 2019, YouTube mistakenly purged numerous accounts and videos associated with cryptocurrency education [Danny Nelson, 2019].

Another approach is to condition the size of the fine on the user’s type. For instance, failing to penalize a popular user results in a larger fine for the platform. Such type-dependent regulation has the potential to achieve the first best outcome ex post. However, this method also presents challenges. First, since the platform cannot pay infinitely large fines, the regulator needs to possess advanced enough technology to increase the expected fine. Achieving the necessary technological proficiency might be prohibitively expensive. For tasks like text reading where AI can be utilized, the marginal cost of technological enhancement might be relatively low. In contrast, especially for nuanced or audio-visual content, the marginal cost may remain high. Second, the regulator would need to anticipate the harm caused by violating content for each user type and adjust the fine accordingly, which could be challenging and potentially too costly.

Hence, I propose a mechanism that allows for user-type differentiation but only requires that the relative types of users be observed by the regulator. This approach effectively filters out cherry-picking. Thus, the regulator can impose a different fine depending on whether the platform retains the content of lower or higher type violators, or both. Importantly, the fine is steeper if the content of only higher type violators is retained, compared to retaining content from both types. The reason is that both the violating content and cherry-picking, which in this case means the concealment of high-type violating content, decrease social welfare.

The remainder of the paper is structured as follows: Section 2 introduces the related lit-

---

7The social value takes into account the societal harm a violation causes and the cost of detection.
erature to this study. Section 3 presents empirical evidence supporting the concept of cherry-picking. The baseline model is presented in Section 4, while Section 5 explores its extensions. Finally, Section 6 concludes.

2 Related Literature

This paper contributes to the existing literature on online platform governance as well as the law and economics of crime and torts. Additionally, the model introduced in this paper falls within the scope of principal-supervisor-agent type models.

Online platforms

The economic effects of introducing liability of online platforms were first analyzed in non-formalized studies Buiten et al. [2020], Lefouili and Madio [2022]. My paper provides a formalized model building on some of the economic insights also mentioned in the non-formalized studies, such as the notion of cherry-picking. The closest models to mine are developed by Liu et al. [2022] and Madio and Quinn [2023]. They study platforms’ incentives to engage in content moderation under the two prevalent business model for revenue source, advertising and subscription. The main overlapping feature of their model with mine is that the platform uses content moderation to increase revenue and may not have incentives to improve the technology/moderate (some) users. In other words, content moderation does not merely depend on technological capabilities but also on economic incentives. The main novelty of my paper is to look at how regulations change economic incentives of the platform and hence the policing strategy the platform implements.

Jiménez Durán [2022] studies the incentives of social media platforms to enforce regulations and ban users for toxic content. The platform monetizes users’ participation with ads and trades-off users’ engagement from both safe and unsafe content. As a result, the platform moderates toxic content to the extent to which it raises advertising revenues. The trade-off is similar to the one presented in this paper, however I focus on endogenizing the regulation and allow the platform to strategically react to it.

Wu [2022] examines two channels used by content creators to disseminate unsafe content: a well-known channel that is easier for regulators to monitor, and a secret channel accessible only to sophisticated consumers and challenging for regulators to oversee. The study reveals that increased regulatory stringency, such as censorship, may inadvertently shift communication from the open to the secret channel, potentially causing greater harm. The key distinction in my paper is the endogenization of regulation, while also considering the platform as a strategic
player. Additionally, my model offers a more generic interpretation of social welfare, enabling a more comprehensive analysis of the regulator’s choice.

A larger body of literature focuses on copyright violations and liability policies, more specifically how liability regimes change the incentives of the platform for monitoring and eventually the content available on the platform [Beard et al., 2018, De Chiara et al., 2021, Landes and Lichtman, 2003]. Notably, Jeon et al. [2022] considers a tradeoff similar to the one in this paper: low-quality merchants (including intellectual property rights infringers) might increase the platform’s sales and hence revenue, and therefore, the platform may not have incentives to monitor and delist the IP infringer merchant. The primary distinctions in my paper are the platform’s strategic removal of individual users’ content and the assumption that copyright violations are not expected to result in significant social harms.

Law and economics literature of crime and liability

Becker [1968] starts the economics literature of crime with the model in which the offender realizes a gain from the violation. In my model, however, (part of) the gain is not realized by the offender but by the platform. The main difference stems from the detection incentives and technology. The platform has an incentive to cover the violation of the user, hence, the technology is not only endogenous to the regulator’s choices but also to the platform’s choice. That is, the platform is not a benevolent public enforcer.

The platform acts as both the beneficiary of the user’s output but also as the private enforcer of the law. Becker and Stigler [1974] argue for private enforcement over public enforcement, especially on the basis of aligned incentives and cost effectiveness. The idea of this paper challenges this view, as the private enforcer’s incentives may not be aligned to social the optimal one, moreover, the private enforcer is subject to different constraints. Polinsky [1980] compares private and public enforcement of fines. However, his model does not involve the main friction of this paper, that is the private party may not want to enforce the law.

The paper also relates to the vicarious liability models started by Sykes [Sykes, 1981, 1984, 1988] and Kornhauser [1982]. The main difference, however, is that the contract between the platform and the user is standardized, the platform cannot internalize the regulation in the contract. The closest paper to mine that models vicarious liability is Arlen and Kraakman [1997], in particular the extension of adjusted quasi-strict liability. Adjusted quasi-strict liability is a strict liability regime with the sanction the firm, or in this case, the platform, faces varying depending on whether the firm reports the wrongdoing of the user. The regime implemented in my model is also an adjusted quasi-strict liability regime, with the fine being 0 if the platform reports/removes the violations itself.
Principal-supervisor-agent models

The model also relates to the literature on principal-supervisor-agent (PSA) theory that was started by Tirole [1986] with the idea of including a supervisor in the standard principal-agent model. The supervisor’s role is to obtain more information about the agent’s activity than the principal could do. The emphasis of the PSA literature is on the potential collusion between parties, especially between the supervisor and the agent, and what the optimal contract the principal offers is. This paper’s model can be interpreted as a special case of such models where the principal is the regulator, the supervisor is the platform, and the supervisor and the agent may collude. Note that one crucial difference between my model and PSA is that the supervisor does not directly work for the principal, therefore he cannot be offered a contract that could incentivize him for truth-telling. The regulators only have punishment (fine) under their disposal for disciplining, but no rewards.

Tirole and Laffont [1991] and Suzuki [2018] analyze a regulatory framework where the agent (firm) and supervisor (government or agency) can collude and falsify evidence reported to the principal (supranational regulator). In this setup, the regulator observes only an aggregate measure, such as air quality in Suzuki [2018], but not the specific type of the agent, which is similar to the approach in my model. The government may observe the firm’s type and report it to the regulator. These papers investigate the optimal contracts between the regulator and government, and between the regulator and firm, in scenarios where collusion (side contracts between government and firm) is either impossible or possible. To ensure truthful reporting, the principal’s problem includes an incentive constraint for the supervisor. My paper can be seen as addressing the scenario where this incentive constraint cannot be satisfied.

2.1 Empirical evidence for cherry-picking

2.1.1 Twitch’s “Do-not-ban list”

In October 2021, an unprecedented leak of Twitch data occurred, revealing a list of streamers with the file name "do-not-ban-list" [Grayson, 2021]. This list suggested that top streamers were given special treatment when it came to suspending their streaming accounts. According to anonymous employees, Twitch’s punishment mechanism worked as follows. Twitch administrators received reports against streamers who were believed to have broken the rules. Punishments for non-partnered streamers were decided by the administrators, whereas matters involving partnered streamers were escalated to a separate partner conduct team, which would issue warnings or make its own judgments [Grayson, 2021]. The "do-not-ban-list" served as a reminder to administrators that they could not decide on punishing certain streamers them-
selves and instead had to escalate the issue to other Twitch staffers or make exceptions to the rules.

Although the exact workings of Twitch’s sanctioning mechanism were not revealed, this example illustrates the concept of cherry-picking. Top users on the platform (partners) were given exemptions when they violated the platform’s rules because their contribution to the platform’s revenue was deemed too important to just have them banned. Consequently, different staff members made decisions about their potential sanctions based on guidelines that may have been different from those used for other users. Non-partnered users, on the other hand, could have been punished more quickly and with less contemplation by the administrators.

2.1.2 Youtube

Take-down policy

Erickson and Kretschmer [2018] explores the factors that prompt copyright owners to request the removal of potentially infringing user-generated content. These measures, known as “notice-and-takedown”, are available in the European Union through the Directive on Electronic Commerce (2000/31/EC) and under Section 512 of the U.S. Copyright Act. According to their findings, there is an inverse relationship between the number of views of a video and the likelihood that YouTube will remove it. Since views increase the revenue of the platform, the incentives of YouTube are similar to the platform’s described in the model.

Youtube Partnership Program

In 2018, YouTube faced a significant crisis known as the "Adpocalypse," where a massive advertiser boycott highlighted the platform’s content moderation challenges Nicas [2017], Alexander [2019]. As a response to this backlash, YouTube tightened the eligibility criteria for its partnership program. This adjustment led to a reduced number of users being able to monetize their content. As a result, videos without monetization were often sidelined by YouTube’s algorithm in favor of content that generated more revenue for the platform Kumar [2019]. Additionally, the updated partnership program introduced a threshold for users to contest YouTube’s monetization decisions, a move that disproportionately benefited larger users by giving them a unique privilege to appeal Patel Sahil [2017].
3  Model

3.1  Model Description

The model considers a social media platform with a large number of users, such as YouTube or Twitter. Users generate content for the platform. Each user (she), can choose between two actions when contributing to the platform: she can either comply with the rules set by policymakers (denoted by $C$) or violate them (denoted by $V$). Each user has a type $x$, representing her popularity, such as the number of followers or subscribers. The user population has a probability density function $g(x)$, a cumulative density function $G(x)$, and a support range $x \in [0, 1]$. Additionally, the pdf satisfies $g'(x) \leq 0$ that is there are weakly fewer higher type users than lower type ones. The continuous support captures the vast number of users on these platforms. A user with type $x$ generates revenue for the platform $I(x, A)$ where the action $A$ can be either compliance or violation. Content that violates the rules results in harm, $h_p(x, V) > 0$, to the platform. However, the net profit (revenue minus harm) from such content, represented by $r(x, V) = I(x, V) - h_p(x, V)$, can be either positive or negative. For tractability, I assume that $r(x, V)$ is either monotonically increasing or monotonically decreasing based on the user’s type across the entire domain:

$$\frac{\partial r(x, V)}{\partial x} > 0 \quad \text{or} \quad \frac{\partial r(x, V)}{\partial x} < 0$$

Violations cause societal harm. This harm is a result of users breaching legal standards or societal norms. Recent studies, such as Jiménez Durán et al. [2022], Müller and Schwarz [2021, 2023], Olteanu et al. [2018], Bursztyn et al. [2019], have underscored the connection between online behavior on social media platforms and real-world criminal activities. Neither the platform nor the user internalizes the social harm caused by violations. A notable instance of online actions leading to offline consequences is the events at Capitol Hill Andrews [2020].

This framework is reminiscent of classical liability models introduced by Becker [1962] and later expanded to encompass vicarious liability in works such as Polinsky [1980] and Arlen and Kraakman [1997]. In these models, as in this paper, agents’ violations increase a principal’s instantaneous income. A unique aspect of this model is the differentiation of users by their

---

8According to the Digital Services Act, illegal content is defined as "any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law." (Article 3h)

9The platform’s revenue may come from advertising or subscription fees; the model does not differentiate between them.

10Both YouTube and Twitch have both types of income sources.
type, \( x \), which determines revenue, rather than by wealth. Social welfare is represented by \( W(x, A) \) with \( W(x, V) = I(x, V) - h_p(x, V) + h_s(x) \) and \( W(x, C) = I(x, C) - h_p(x, C) + h_s(x) \). Here, \( h_s \) denotes the societal value of the content, encompassing societal harm, detection, and enforcement costs, as well as benefits like disseminating accurate news or educational material. The paper primarily concentrates on harmful content, hence generally focusing on \( h_s(x) < 0 \).

Importantly, while the platform might favor a violation as the user’s action if \( r(x, V) > r(x, C) \), it could be the inefficient action if \( W(x, V) < W(x, C) \). When the platform deletes a user’s content, both the associated harm and revenue are negated.

The regulator monitors the actions of both the user and the platform. If violating content is detected on the platform, the platform can be subject to a monetary fine \( f < \tilde{f} \). However, if the platform correctly addresses the violation or if no violation took place, no fine is imposed. The fine is set endogenously by the regulator, however, it cannot exceed \( \tilde{f} \), the maximum fine the platform can pay. The regulator has a technological disadvantage in two dimensions. First, the regulator cannot monitor all content. Hence, violations can be identified by regulators with probability \( p \), and I explore both exogenous and endogenous technology in my analysis in subsequent sections. Second, the regulator may be unable or unwilling to differentiate based on the users’ type \( x \). This assumption aligns with the already enacted regulation of the NetzDG since sanctions are not dependent on the type of the user. In contrast, the platform can observe all the actions and the types of the users costlessly.

A key assumption of this model is that the regulator’s sanctions are directed solely at the platform, not individual users. This approach aligns with prevailing regulatory practices. The rationale behind this regulatory choice, although not explicitly stated by authorities, could stem from several factors. For instance, users may be spread across different countries and jurisdictions, complicating the identification of the appropriate regulatory body for enforcement. Furthermore, users often lack the financial means to pay substantial fines, making enforcement potentially more costly than the fines collected.

The actions of the game are illustrated in the figure below.

---

11 This assumption is relaxed in later sections.
12 Section 6.3 examines scenarios where users are also subject to liability.
The timeline of the model is as follows.

$t=0$: Nature draws each users’ type $x$ which is privately observed by both the user and the platform.

t=$1$: The regulator sets the fine, $f$.

t=$2$: The users take action, they either violate or comply.

t=$3$: The platform observes the action of the users and decide which content to retain and which to delete.

t=$4$: The regulator samples from the platform and imposes a fine if a violation is detected.

### 3.2 Comments on the model setup

#### 3.2.1 NetzDG

The model setup resembles the legislative framework of the German Netzwerkdurchsetzungsgesetz (NetzDG), which was introduced in 2018 to combat hate speech on major social media platforms. Similar to the model, the rules of NetzDG involve imposing a penalty fee on platforms, regardless of the user’s type, if violations are detected on the platform that the platform fails to remove appropriately and promptly.\(^{13}\) However, one key difference between the NetzDG

---

\(^{13}\)The penalty fee increases proportionally with the size of the social network platform size, as measured by the number of users, but not by the size of the individual user. Moreover, all the largest social media networks fall into the same category.
and the model is that in the model, users are not held liable for violations, whereas they may face a penalty fee under the NetzDG. This distinction is relevant in an international context, as individual liability may be difficult to enforce and costly, while individuals may lack the financial resources to pay monetary fines. Hence, the NetzDG stipulates that platforms may be subject to much larger fines than individuals. Individual liability is considered in section 6.3.

The enforcement mechanism of the NetzDG is not specified in the regulation. Users can report content they believe to be unlawful, and the platform has the discretion to delete or take no action on the reported content. If the regulator detects violations that are not punished by the platform, both the user and the platform may be held liable and subjected to a preset penalty fee. Unlike in the model, the NetzDG imposes fines on platforms for systemic failures rather than individual cases.\textsuperscript{14} and platforms may be subject to much larger fines than individual users. However, the exact mechanism by which user reports lead to sanctions on the platform is unclear.

3.2.2 Digital Services Act

The Digital Services Act of the European Union (DSA) introduces liability provisions for online platforms, primarily targeting large platforms, regarding third-party content, coming in full force by the 17th of February, 2024. Liability of individual users is not the main focus of the DSA. According to Article 3(h), illegal content is defined as any information that does not comply with Union law or the law of any Member State that aligns with Union law, regardless of the specific subject matter or nature of that law. Under the DSA, platforms have the option of conducting voluntary content moderation (Article 7), where they independently investigate and take measures to detect, identify, remove, or disable access to illegal content. Consequently, platforms can be exempt from liability for hosting illegal activities or content if they promptly remove or disable access to such content (Recital 22). The model primarily emphasizes that platforms can avoid liability by engaging in voluntary content moderation, driven by the potential liability they may face. However, it is important to note that the Digital Services Act (DSA) includes additional measures to regulate online platforms. These measures encompass mandatory content moderation, the involvement of trusted flaggers, and independent auditing.

**Mandatory content moderation** The DSA establishes mandatory content moderation (Article 9), wherein platforms are required to take action against illegal content upon receiving an administrative or judicial order \cite{Hoboken et al., 2023}. In this scenario, mandatory content...
moderation can be seen as a mechanism for regulators to observe the platform’s actions regarding violating content directly. This is a less interesting part of the moderation policy, since the platform would comply by removing the illegal content in equilibrium to avoid having to pay the fine. Mandatory content moderation leaves out the part of the implementation on how the regulator may uncover the violations.

**Trusted flaggers** Articles 16 and 22 of the DSA establish that platforms must provide a mechanism for individuals or entities to notify them of potentially illegal content. In particular, notifications form the so called trusted flaggers must be dealt with without delay. One advantage of the system is that it allows multiple individuals to report the same content which could indicate that the content may be illegal with a higher likelihood. However, content created by high type users tends to reach a larger audience, increasing the likelihood of it being reported. As a result, individual reporting ultimately relies on the regulator’s ability to monitor and evaluate the platform’s decisions regarding reported content. Consequently, the efficacy of individual reporting hinges on the regulator’s capacity to monitor and assess the platform’s responses to reported content, a facet embedded within the model’s framework.

**Independent audit** Article 37 of the DSA stipulates that providers of very large online platforms and search engines are required to undergo independent audits, conducted at their own expense and at least once a year. In the model’s context, this means that a private auditor replaces the regulator in assessing the platform’s compliance. Hence the model remain unchanged if the auditor has the same incentives as the regulator.

In summary, the mechanisms discussed ultimately center around the regulator’s ability to monitor the platform’s actions, an aspect highlighted in the model. Whether through voluntary or mandatory content moderation, involvement of (trusted) flaggers, or implementation of independent audits, the regulator has a pivotal role in overseeing and ensuring compliance with regulations.

---

15 Another advantage of the system of trusted flaggers is that the knowledge of experts is elicited in deciding what content is illegal. However, I do not explicitly model this in the model rather assume the regulator is the best expert.

16 Arguably, the private auditor may not fully endogenize the external harm, hence the outcome may be worse from a social point of view than in the model.
4 Analysis

The analysis proceeds in several steps. First, I present a baseline scenario where the platform does not purposefully influence the probability that the regulator detects violating content on the platform. In other words, the detection probability is only a function of the regulator’s actions. Subsequently, I allow the platform to influence detection by strategically choosing which content to retain and which to delete, engaging in cherry-picking. I consider both naïve users who have an intrinsic motivation to violate the rules and also strategic users who anticipate the action of the platform correctly with respect to the retention/deletion of their content.

4.1 Baseline model

In this scenario, the regulator varies a fine, \( f_p \), to incentivize the platform to delete violating content. Following the principles established by Becker [1962], increasing the fine, \( f_p \), is recognized as a more cost-effective strategy than enhancing detection technology \( p \). This is because raising the fine incurs no additional costs, whereas improving technology demands costly investments. Consequently, I keep the detection probability, \( p \), constant in this analysis and only vary the fine, \( f_p \).

For simplicity I assume that all users, irrespective of their type \( x \), violate the rules, similar to Wu [2024]. Users derive no utility from these violations and face no liability for them. This assumption can be motivated by the fact that users may not consider how a policy change by a regulator alters the platform’s incentives for content moderation, thus they are naïve. This assumption does not change the message of the baseline model and is relaxed later.

4.1.1 Deterrence

The payoff of the platform from retaining the content of a user with type \( x \) is

\[
\pi(x, V, R, f_p) = I(x, V) - h_p(x) - pf_p, \tag{2}
\]

where \( R \) denotes retaining the content. To simplify notation, I use \( r(x, V) = I(x, V) - h_p(x) \), as the net income of the platform.

\[
\pi(x, V, R, f_p) = r(x, V) - pf_p. \tag{3}
\]

If the platform deletes the content, denoted by \( D \), the payoff is 0.
\[ \pi(x, V, D, f_p) = 0 \] (4)

Hence, the platform retains content if \( r(x, V) - pf_p > 0 \). The platform’s overall profit is

\[ \pi(x, V, R, f_p) = \int_V r(x, V)g(x)dx - p \int_V f_p g(x)dx. \] (5)

The regulator does not make the fine, \( f_p \) a function of the type of the user. The fine does not enter the welfare function directly since that is a transfer between the platform and the regulator. Denote the welfare function with \( W \).

\[ W(x, V, f_p) = \int_V r(x, V) + h_s(x)g(x)dx. \] (6)

Consider that for some \( x \), \( r(x, V) > 0 \) and \( W = r(x, V) + h_s(x) < 0 \) holds, meaning the platform benefits from retaining the content of some users on the platform, however, the content is harmful, \( h_s(x) < 0 \) with \( r(x, V) < -h_s(x) \), once societal harm is considered. For example, such content, like fake news tweets, engages readers and increases their time on the platform, allowing for more advertisement exposure and boosting the platform’s revenue. However, this content causes societal harm due to the misinformation it spreads [Van Alstyne et al., 2023, Bhargava, 2022].

The regulator’s objective is to set the fine so that it incentivizes the platform to remove content if it results in societal harm, \( W(x) < 0 \), and to retain content if it contributes positively to societal welfare, \( W(x) > 0 \). Moving forward, I differentiate cases based on the interplay between the platform’s optimal decisions and the regulator’s preferences, as determined by the functions \( r(x, V) \) and \( h_s(x) \). I assume that these two functions intersect at a single point. No intersection yields a trivial result, while multiple intersection can be fundamentally reduced to a combination of the subsequent two cases discussed.

**Lemma 4.1.** If a violation of low types is efficient, the regulator either imposes the maximum fine or does not impose any fine.

Proofs are in the Appendix.

Figure 2 depicts the above condition visually, with \( V_s \) and \( C_s \) denoting the socially efficient action. Hence, \( x_s \) becomes the socially efficient cutoff. In words, violating is a socially efficient action for low types, however, it becomes socially inefficient for the high types. A possible interpretation is that the social harm increases as the type of the user increases, outweighing the benefits, while the detection cost is constant, making it inefficient to detect violations of the
lowest types. Note that the detection costs are embedded in \( h_s(x) \). Since the platform’s profit is increasing in type, large enough fines are required to incentivize the platform to delete the content of higher types. However, a large fine also leads the platform to delete socially efficient content. If the regulator does not impose a fine, all users commit violations and the platform does not delete any of them. Under the maximum fine, the platform deletes all content, meaning the platform ceases to exist. Which extreme is optimal depends on how harmful the content is without any liability when all users are violating, that is, whether

\[
W(x, V, 0) > W(x, V, \bar{I}) \tag{7}
\]

**Lemma 4.2.** If violation of high types is efficient, the first best outcome is achievable.

\[
f_p^* = \frac{r(x_s, V) + h_s(x_s)}{p} \tag{8}
\]

Below graphic visually shows the condition.

![Figure 2: Violation of low types is efficient](image)

The fine is not type dependent, yet it induces the efficient outcome. The reason is that the fine needed to ensure the platform has sufficient incentives to punish type \( x_s \) is also large enough to deter lower types. This fine exceeds the revenue minus the harm due to the imperfect technology \( p \). Furthermore, the state lacks the incentive to enhance the technology if it can simply increase the fine, a notion consistent with the early findings of Becker [1968].

### 4.2 Cherry-picking

In this section, the platform responds to the regulator’s enforcement strategy by strategically deleting content from users on the platform. This approach is adopted to reduce the likelihood of incurring fines. Consequently, the effectiveness of enforcement hinges on the actions of both the regulator and the platform. The regulator conducts random checks on content published on the platform, which may or may not be available at the time of review, and imposes fines for any violating content that the platform has retained. Additionally, the regulator examines content that the platform has already removed to ensure that benign content is not being unjustly deleted.

The regulator’s strategy is characterized by \((n, f_p)\), where \( n \) denotes the number of users
sampled, and \( f_p(\cdot) \) represents the fine structure. In the simplest scenario, the regulator samples once and imposes a fine if a violation is found in content that the platform failed to delete. This approach, which I term 'uniform regulation,' is akin to the policy outlined in Germany’s NetzDG. Proceeding with the assumption that all users are violators, the probability of the platform receiving a fine is determined by the ratio of the mass of retained content to the total mass of users. Therefore, the likelihood of the platform being fined is

\[
p = \frac{\text{mass of violators}}{\text{mass of participants}} = \int_V g(x) dx
\]

that is the integral over the area(s) of violations, that is where the platform does not delete the violating content.

The profit of the platform for any arbitrary \( f_p \) is

\[
\pi(x, f_p) = \int_V r(x, V) g(x) dx - \int_V g(x) f_p d(x)
\]

**Definition 4.3. Cherry-picking:** For violating users, if \( x' > x \), the platform’s action of deleting the content of user \( x \) while retaining \( x' \) is termed as cherry-picking.

By cherry-picking, the platform selectively deletes content from users whose contribution to its profit is low, thereby reducing the likelihood of the regulator imposing a fine. This strategy is feasible because the regulator faces a technological disadvantage in preemptively identifying the users’ types. Essentially, cherry-picking results in preferential treatment for higher type users, allowing them to commit violations that lower type users would be penalized for.

**Proposition 4.4.** Uniform regulation with \( f_p > 0 \) provides incentives for cherry-picking under Case I, where violations of high type users are socially inefficient (figure 3).

Since the regulator’s fine is not conditioned on the user’s type, the platform is incentivized to penalize users who contribute little to its profit, thereby reducing the chance of receiving a fine. However, the platform retains the content of top users to maximize revenue. The platform, however, does not punish the top users to increase revenue. The higher the fine, \( f_p \), the more incentive the platform has to decrease the mass of users who violate.

Next, consider how the regulator sets the optimal fine, \( f_p^* \), and how the platform reacts in terms of the cutoff type \( x_p^* \), that is the lowest type the platform keeps the content of.

**Proposition 4.5.** If the violation is socially optimal for high types,

\( f_p^* = r(x_s, V) \) and \( x_p^* = x_s \) are the optimal choices by the regulator and platform respectively, that is the first best outcome is achieved.
The proof is in the appendix.

If violations by high type users are socially optimal, the first-best outcome can be attained by setting the fine as high as the platform’s revenue at cutoff value $x_s$, or $f^*_p = r(x_s, V)$. This fine incentivizes the platform to remove content from users of type $x < x_s$ while retaining content from those of type $x > x_s$, achieving the first best outcome.

**Proposition 4.6.** If the violation is socially optimal for low types,

\[
\begin{cases}
    f^*_p = 0 \text{ and } x^*_p = 0, & \text{if } W(x, V, 0) > W(x, V, f_p) \\
    f^*_p = \tilde{f}_p \text{ and } x^*_p = r^{-1}(\tilde{f}), & \text{otherwise}
\end{cases}
\]

are the optimal choices by the regulator and platform respectively, and the first best outcome cannot be achieved.

When violations by top users are socially harmful, a uniform regulation where $\tilde{f} > f_p > 0$ cannot be optimal because of cherry-picking. In order to minimize the probability of the fine $f_p$, the platform deletes content from users at the lower end of the type distribution, signaling “good” behavior, consistent with the good Samaritan approach highlighted by [Buiten et al. 2020]. Such deletions reduce overall social welfare. Hence, the classical results of [Becker 1968] that the regulator should increase the fine to improve social welfare does not hold in the current setup.

If violations by high-type users are efficient, setting the fine according to $f^*_p = r(x_s, V)$ solves the regulator’s problem. Therefore, the subsequent sections of this paper focus on the case where violations by top users lead to significant, and often offline, social harms, rendering them inefficient.

### 4.2.1 Strategic users

This section allows for strategic behavior by users (referred to as ‘she’). Specifically, users decide whether to comply or violate rules based on their expected payoff. Under the assumption that violating content generates more revenue than benign content, conditional on being retained, the user’s utility satisfies the following inequality:

\[
U(x, V, R) > U(x, C, R)
\]  \hspace{1cm} (11)

If the user’s content is deleted, her payoff is zero, i.e.,

\[
U(x, A, D) = 0.
\]  \hspace{1cm} (12)
Consequently, the user receives the highest payoff if she violates and her content is retained. However, she prefers retained benign content over deleted violating content. The augmented game’s timing is as follows:

1. All players learn the type-distribution of the users on the platform.
2. The regulator publicly announces the fine $f_p$ that the platform has to pay if violating content is uncovered.
3. Users sequentially choose whether they comply or violate the rules in the order of their type. The largest user moves first, then the second largest etc.
4. Platform decides which content to retain
5. Regulator samples randomly and imposes the fine $f_p$ if violating content is found on the platform.

Users observe the fine the platform may face and also the type of other users on the platform, hence they can calculate the payoff of all users and the platform. 

**Lemma 4.7.** Users with high enough type, $(x \geq r^{-1}(f_p, V))$, violate the rules while the rest comply.

Intuitively, users expect the platform to retain content from top users and delete content that violates rules from those at the lower end of the distribution. Anticipating the deletion of their content, users prefer to comply with the rules. Users thus self-sort into two groups: those who comply and those who violate rules but expect the platform to retain their content regardless. For the equilibrium characterization, denote with $x_u$ the lowest type that violates.

**Proposition 4.8.** In equilibrium, the optimal policy, $f_p^*$ and the optimal cutoff value $x_u^*$ are as follows

$$\begin{cases} f_p^* = 0 \text{ and } x_u^* = 0, & \text{if } W(x, V, 0) > W(x, V, f_p) \\
 f_p^* = \tilde{f}_p \text{ and } x_u^* = r^{-1}(\tilde{f}), & \text{otherwise} \end{cases}$$

Moreover, the platform retains all content.

The proof follows directly from proposition 4.6 and lemma 4.7. Comparing the equilibrium outcome to proposition 4.6, the same set of users’ violations are retained. This is because both

---

17 Alternatively, the user might not derive benefit from compliant content but dislikes having her content removed. This scenario results in a qualitatively similar equilibrium.
18 The setup is similar to [Halac et al., 2020] in which investors of different endowments observe the interest rates ex ante and may opt to invest in a risky project.
the users and the platform have similar incentives: they prefer violations from top users while
disfavoring violations from the bottom. As a result, strategic users at the bottom do not commit
violations and rather opt to publish benign content.

I continue the analysis under the assumption that users are not strategic for two reasons.
Firstly, the information structure required to include them as strategic players is complex and
unrealistic. Secondly, even if users were strategic, the equilibrium outcome would be similar, as
demonstrated above.

4.3 Improving technology

The regulator’s technology is characterized by the size of the sample, \( n \), used for monitoring
whether the platform deletes violating content. Therefore, enhancing this technology can be
achieved by increasing the sample size. However, an increase in sample size does not necessarily
lead to a higher probability of imposing fines due to the platform’s strategy of cherry-picking. To
illustrate this point, consider a scenario where the regulator samples \( n \) times with replacement
and imposes a fine on the platform if any violation is detected\(^{19}\) For simplicity, also consider
that sampling is not costly.

\[
\pi(x, f_p) = \int_V r(x, V)g(x)dx - n\int_V g(x)f_p d(x). \quad (13)
\]

**Lemma 4.9.** Improving the technology, that is increasing the sample size \( n \), provides incentives
for cherry-picking.

The lemma echoes the findings of 4.4, illustrating a consistent pattern: as the regulator at-
ttempts to heighten the expected sanctions, the platform counteracts by intensifying its cherry-
picking strategy. Consequently, enhancing the monitoring technology does not necessarily trans-
late into a higher expected sanction for the platform.

5 Extending the regulator’s strategy

To effectively counteract cherry-picking, the regulator must consider strategies beyond the ran-
dom sampling from all user-generated content and the use of a uniform fine. This necessitates
utilizing information about the types of users on the platform. While specific regulations like the
NetzDG or the Digital Services Act do not explicitly mandate this strategy, its implementation
appears necessary. The following subsection, 5.1 investigates the scenario where the regulator
assigns in the sampling process, based on their types. Meanwhile, subsection 5.2 moves away

---

\(^{19}\)Given the continuum of users, the choice of sampling with or without replacement does not alter the outcome.
from a uniform fine structure, introducing fines that vary depending on user type. Note that the solutions are not directly comparable since the regulator’s choice set differs.

5.1 Weighting users

It is plausible that the regulator possesses ex ante knowledge about the types of users, for example, based on the number of followers or subscribers a user has. Anticipating the platform’s cherry-picking strategy, the regulator might design rules that place additional emphasis on auditing top users. Instead of selecting randomly from the user base, the regulator could implement an ex ante announced rule of selection, denoted as $w(x)$. This changes the probability of detection to:

$$p = \int_V w(x)dx,$$  \hspace{1cm} (14)

Here, the regulator’s preferred distribution $w(x)$ replaces the true population $g(x)$. The firm’s profit then becomes:

$$\pi(x, f_p) = \int_V r(x, V)g(x)dx - \int_V w(x)f_p d(x),$$  \hspace{1cm} (15)

The regulator’s objective is:

$$W(x, V, C, f_p) = \int_0^1 r(x, V) + h_s(x)w(x)dx > 0,$$  \hspace{1cm} (16)

In this scenario, the regulator has the ability to choose the function $w(x)$. By choosing a type-dependent weight, the regulator can achieve the first best in numerous ways. For example, consider the rule that the regulator only samples if the user has a large enough type, that is from $x > x_s$. The probability of a fine then becomes:

$$p = \int_{V|x>x_s} g(x)dx,$$  \hspace{1cm} (17)

The firm’s profit under this regulation is:

$$\pi(x, f_p) = \int_V r(x, V)g(x)dx - f_p \int_{V|x>x_s} g(x)dx,$$  \hspace{1cm} (18)

Since the regulator never chooses users with $x < x_s$, the platform does not penalize these users, which is the efficient action. Here, cherry-picking means the platform penalizes users when it is efficient to do so. However, unless the weights are sufficiently type-dependent, the regulator might still fail to penalize top users. A further challenge with this regulatory approach
is the necessity to announce the weights ex ante to guide the platform’s actions. Consequently, both the platform and users are aware of which users have a diminished (or potentially zero) chance of selection, which the platform could further exploit by overregulating, or individuals could oppose such regulation when the regulator announces that the decision on their content is not being monitored. Therefore, while weighting users might achieve the first best outcome, its implementation by a regulator seems improbable.

The model assumes that all illegal content is reported, as platforms are not liable until a third party flags the content. In reality, the frequency of reporting can differ; notably, content from more popular users may attract more reports. Nonetheless, under current regulations like NetzDG and DSA, a single report obligates the platform to respond. Since the model focuses on user content that generates revenue, it presumes popular enough users. Therefore, the assumption that the probability of at least one report of violating content is equal across types looks realistic.

5.2 Type dependent fine

The standard solution for optimal deterrence, provided by Becker [1968], is to set the fine high enough so that

\[ f_p(x) = -\frac{h_s(x)}{p}. \]

Equation (19)

In this case, the size of the fine depends on the type of the user, for example, on how many followers the user has or how many views the content has received.

The probability \( p \) represents the likelihood that the regulator audits a specific user’s content. Given that the type distribution is modeled as a continuous variable, we have \( \Pr(X = x) = 0 \). Hence, consider that the regulator pools types together to create a discrete distribution with groups \( x_1, x_2, \ldots, x_n \). For instance, \( x_1 \) might represent the lowest percentile in terms of follower count, while \( x_{100} \) represents the highest. The optimal fine, \( 20 \), can then be rewritten as

\[ f_p^*(x_i) = -\frac{h_s(x_i)}{p(x_i)}. \]

Equation (20)

where \( i \) belongs to the set \( \{1, 2, 3, \ldots, n\} \). Given the potential for numerous groups \( x_i \), the implementation of such regulation can closely approximate the first best outcome. However, this approach presents potential challenges. First, considering that the platform’s maximum payable fine is \( \bar{f} \), achieving the necessary technological proficiency might be prohibitively expensive. For applications like text reading where AI can be employed, the marginal cost of technological enhancement might be relatively low. In other contexts, especially when the con-
tent is nuanced or audio-visual, the marginal cost may remain high. As highlighted in previous
sections, expanding the range $x_i$ for a given fine $f_p(x_i)$ can lead to increased cherry-picking by
the platform. Second, the regulator would need to anticipate the harm caused by violating con-
tent for each type $x$ and adjust the fine accordingly, which could be challenging and potentially
too costly.

One feature that a type-dependent regulation misses is that it does not explicitly anticipate
the cherry-picking of the platform. The next section introduces a mechanism that incorporates
such a feature.

5.3 Fines as a function of relative type

This section proposes a mechanism to mitigate cherry-picking by combining enhanced tech-
nology with the notion that the regulator possesses some information about user types on the
platform. The enhanced technology allows the regulator to sample from the platform multiple
times. To maintain a realistic and limited information requirement, consider the following
mechanism: The regulator samples twice from the platform and observes not only the users’
actions but also their relative types, i.e., which user’s type is higher. This data might reflect
the user’s network size, indicated by metrics like the number of followers, subscribers, or views.
Such information is publicly available on popular platforms like YouTube or Twitch. Essen-
tially, the regulator aims to influence the platform’s behavior to retain content from low-type
users and delete content from high-type users, aligning with socially efficient actions, under the
assumption that all users violate.

Under this information structure, the regulator can impose the following fines:

- $f_1$: the fine if the platform deletes the content of the lower type user but retains that of
  the higher type user, i.e., if cherry-picking is detected.

- $f_2$: the fine if the platform deletes the content of the higher type user but retains that of
  the lower type user.

- $f_3$: the fine if the platform does not penalize either user.

The subsequent graph illustrates full cherry-picking.

\[
\begin{array}{c}
C_{p'} \\
0 \quad x'_{p'} \quad 1 \\
V_{p'}
\end{array}
\]

Figure 3: Full cherry-picking
\( C_{p'} \) denotes the range where the platform deletes the content of the violating user, indicating compliance with the regulator’s rules from the platform’s perspective, while \( V_{p'} \) denotes the area where the platform retains the violating content.

The platform’s profit under full cherry-picking is

\[
\pi_2(x_{p'}) = \int_{x_{p'}}^{1} r(x)g(x)dx - 2G(x_{p'})(1 - G(x_{p'}))f_1 - (1 - G(x_{p'}))^2 f_3. \tag{21}\]

To decrease cherry-picking, the area with violations at the top, represented by \( V_{p'} \), must be reduced. Consider the alternative strategy depicted in graph 4.

Decreasing cherry-picking means reducing the area with violations at the top, denoted by \( V_{p'} \). Consider the alternative strategy depicted on graph 4.

\[ \begin{array}{ccc}
V_1 & C_1 & V_2 \\
0 & x_1 & x_2 & 1 \\
\end{array} \]

Figure 4: Decreased cherry-picking

The corresponding profit is

\[
\pi_2(x_1, x_2) = \int_{0}^{x_1} r(x)g(x)dx + \int_{x_2}^{1} r(x)g(x)dx
\]

\[
- [G(x_1) + 1 - G(x_2)]^2 f_3 - 2[G(x_2) - G(x_1)](1 - G(x_2))f_1 - 2G(x_1)[G(x_2) - G(x_1)]f_2
\]

\( (22) \)

The objective function of the regulator is

\[
\max_{h_1, h_2, h_3} W = \int_{0}^{x_p} r(x)g(x)dx + \int_{0}^{x_p} h(x)g(x)dx \tag{23}\]

Lemma 5.1. To alleviate cherry-picking, \( f_1 > f_3 \) must hold.

The fine must be larger if the user with higher type violates. This decreases the platform’s incentive for cherry-picking.

Proposition 5.2. To alleviate cherry picking \( f_1 > f_3 \), must hold.

While it may seem surprising to punish fewer violations with a higher fine, the reasoning is straightforward. From the regulator’s perspective, two actions warrant sanctions. These are the user’s violation and the platform’s attempt to cover it through cherry-picking. Therefore, imposing a large fine when the regulator detects that content from a higher type user is retained while that of a lower type user is not, aims to disincentivize cherry-picking.
Alternatively, if the regulator prefers to stop violations at the top, that is

\[ \begin{array}{ccc}
C_1 & & V_1 & & C_2 \\
0 & x_1 & & & x_2 & 1
\end{array} \]

Figure 5: Decreased cherry-picking

Yields a similar condition on \( f_1, f_2, f_3 \) based on a similar argument.

### 5.4 First best with sufficiently high \( f_1 \)

Proposition 5.2 establishes the importance of imposing a sufficiently high fine \( f_1 \), that explicitly disincentivizes cherry-picking. The larger \( f_1 \) is, the larger is the expected punishment for cherry-picking, therefore the platform engages less in it.

Consider that the maximum fine the regulator can impose, \( f_1 \), is sufficiently high. \(^{20}\)

The regulator’s objective function is

\[
\max_{f_1, f_2, f_3} W = \int_0^{x_p} r(x)g(x)dx + \int_{x_p}^x h_s(x)g(x)dx
\]  

The regulator’s objective to reach the first best outcome can be broken into two parts as follows. First, the platform must have incentives to punish violations by the top users.

The platform’s profit under the behavior preferred by the regulator is

\[
\pi_1(x_p) = \int_0^{x_p} r(x)g(x)dx - G(x_p)^2 f_3 - 2G(x_p)(1 - G(x_p))f_2
\]  

The platform’s profit under cherry-picking, however, is

\[
\pi_2(x'_p) = \int_{x'_p}^1 r(x)g(x)dx - 2G(x'_p)(1 - G(x'_p))f_1 - (1 - G(x'_p))^2 f_3
\]  

Second, the platform has to choose \( x_p = x_s \) for the first best outcome to realize.

**Proposition 5.3.** The optimal policy either reaches first best or there is underdeterrence in equilibrium.

The regulator sets \( f_1 \) high to deter cherry-picking and assigns \( f_2 = 0 \), indicating no penalty if the platform behaves in line with the regulator’s preference. The fine \( f_3 \) for allowing both violations is set to discourage cherry-picking and to guide the platform’s behavior towards the desired outcome. Given that \( f_3 < f_1 \) needs to hold, setting \( f_3 \) too high could counteract the

\(^{20}\) It is common to assume that individuals have a finite wealth, see for example Polinsky and Shavell [1984]. Platforms, however, face astronomical fines in regulations, such as the GDPR.
deterrent effect of $f_1$ on cherry-picking. If $f_3$ is lower than its optimal value, underdeterrence occurs because the platform faces a reduced expected fine. Conversely, over-deterrence is not a concern; if $f_3$ exceeds its optimal value, the regulator can simply reduce it while ensuring $f_3 < f_1$.

6 Model extensions

In this section, I explore extensions to the model, accounting for users’ intrinsic preferences to adhere to regulations and the platform’s potential imperfections in detection technology.

6.1 Compliant agents

This part introduces that some users do not commit violations and comply with the rules. Content that complies with the rules and creates positive value for the platform and for society is efficiently kept on the platform. Therefore, I disregard this type of content. By doing so, the regulator’s job is simplified, in other words, considering these content increases the incentives for cherry-picking. For simplicity, I assume that the net value from compliance to the platform is $r(x, C) = r_C < 0 \ \forall x$, while the net social value is $s_C > 0$. Examples of compliant content being removed include YouTube’s suppression of videos related to the Hong Kong protests in 2019 [Oremus 2019] due to concerns about aggressive content discouraging advertisers. In late 2022, Twitter removed accounts of several journalists at the request of Elon Musk [Sayantani 2022]. Moreover, in 2019, YouTube mistakenly deleted numerous accounts and videos related to cryptocurrency education [Danny Nelson 2019]. In terms of regulation, however, platforms are also not allowed to arbitrarily delete content [Global Freedom Expression 2023].

If the regulator detects content that is not a violation but deleted by the platform, a fine of $f_C$ is imposed on the platform. If violating content is found on the platform, the fine is $f_V$. Consider that $z$ is the fraction of compliant users. Denote with $k$ the ratio of the benign content the platform decides to keep. The platform’s objective is

$$
\pi(x, r_C, r_V) = z(kr_C - (1 - k)f_C) + (1 - z) \int_V r(x)g(x)dx - Pr(f_V)f_V
$$

(27)

Importantly, $Pr(f_V) = \int_{x_p}^1 g(x)dx$, just as before, and hence the same first order condition arises as in (13). The reason why the regulator monitors all content that is produced is now

21 In Europe, social media platforms have to consider fundamental rights, shown for example by the cases Forza Nuova v. Facebook Ireland NTD and Ein Prozent v. Facebook Ireland Ltd [Global Freedom Expression 2023]. In the US, Section 230 gives more opportunities to platforms to delete content they dislike, however, new laws are emerging there too, especially around censoring political candidates. See for example NetChoice v. Attorney General, State of Florida [Global Freedom Expression 2023].
also explicitly shown. The platform has incentives to delete content that is not a violation.

The first order condition with respect to the compliant content is

$$z \left( \frac{\partial \pi(x, r_C, r_V)}{\partial k} \right) = z(r_C + f_C) = 0$$

That is, the platform’s problem has a corner solution. It keeps all the content if $$f_C > -r_C$$, and deletes all otherwise. The regulator hence sets up $$f_C$$ so that the above inequality is satisfied and the platform keeps the benign content on the platform in equilibrium.

### 6.1.1 Policy

The regulator might weigh the harm caused by not removing harmful content against the harm of deleting benign content. Notably, the issue of cherry-picking can emerge when the regulator aims to address both simultaneously. By intending to oversee both existing and previously deleted content, the platform can strategically remove content based on the revenue it produces. Consequently, an alternative approach surfaces where the regulator concentrates solely on currently available content. Through this method, cherry-picking is eradicated, but at the expense of permitting the platform to delete any content they choose, thereby negating its social value.

By not monitoring the content that platforms remove, regulators effectively allow platforms to exclude any material they choose. If platforms foresee potential backlash from subscribers or advertisers regarding specific content, they might opt to remove it, even if it complies with their terms of service. This becomes particularly concerning given that a substantial segment of the population relies on digital platforms for news [Forman-Katz and Matsa 2022](#22).

Furthermore, evidence suggests that platforms might be motivated to censor certain content [Oremus 2019](#27). Another point of contention is the potential lack of incentive for platforms to rectify errors, exemplified by YouTube’s decision to remove educational cryptocurrency videos [Danny Nelson 2019](#).

### 6.2 Platform’s imperfect technology

This section reconsiders the assumption that the platform possesses the perfect detection technology for identifying violations. In practice, regulations like NetzDG account for potential inaccuracies in the platform’s violation detection capabilities. As such, the regulator neither assumes nor demands that the platform’s detection technology be perfect. Here, “technology”

---

22Roughly 50% of Americans source their news from social media, as indicated by [Forman-Katz and Matsa 2022](#22).

27
denotes the platform’s capacity to precisely assess users’ actions in alignment with the regulator’s perspective.

6.2.1 Compliant agents

Consider that the platform’s technology is imperfect, meaning the platform categorizes the users’ content correctly with a probability less than 1. In essence, the platform only receives a signal about the action taken by the user, and the quality of the signal is independent of the user’s type. Identifying a violation by a higher or lower type user is equally challenging for the platform. The efficacy of the platform’s technology does not directly influence how users interpret the content. Rather, the technology serves to identify violations and helps the platform avoid sanctions from the regulator.

If the platform categorizes the user’s content to be a violation, it is truly a violation with probability \( \frac{1}{2} \leq q_V \leq 1 \). Similarly, if the platform categorizes the content as compliance, it is correct with \( \frac{1}{2} \leq q_C \leq 1 \). Consider that the platform’s income from benign content is \( r(x, C) > 0 \) with \( \frac{\partial r(x, C)}{\partial x} \geq 0 \).

Lemma 6.1. The platform cherry-picks, that is the platform does not delete the content if the type of the user is high enough.

Proof follows in the text.

The platform’s choice to retain violating content boils down to the inequality

\[
\pi(x, V, q) = q_V [r(x, V) - pf_V] + (1 - q_V) r(x, C) > 0
\]

Given technology \( q_V \), the platform retains the content of the user if her type is large enough, that is

\[
q_V r(x, V) + (1 - q_V) r(x, C) > pf_V
\]

Since \( r(x, V) > r(x, C) \) \( \forall x \), the inequality is satisfied for high enough types. In particular, if the firm has the worst technology, \( q_V = \frac{1}{2} \), the platform still retains the content of types for which \( r(x, V) - r(x, C) > pf_V \) is satisfied. In other words, if the type of the agent is high enough, the platform is going to retain her content. The better the technology is, that is the more certain the platform is that the user’s content is a violation, the higher the threshold becomes for cherry-picking.

\[23\] Both a constant revenue for the platform from the compliant users’s content, like before, or a revenue function that is increasing in the type yield similar results. Allowing some of the benign content to yield negative revenue for the platform would also leave results qualitatively intact.
Next, the platform is certain with $q_C$ probability, if the user’s content is categorized as compliance. The platform’s profit is then

$$\pi(x, C, q) = q_C r(x, C) + (1 - q_C) \left[ r(x, V) - pf_V \right] > 0$$ (31)

Consider low types such that $r(x, V) - pf_V < 0$ is satisfied. Hence, the platform needs to be certain enough, that is $q_C$ must be high enough for the platform to retain the content. Similarly to the violation case, if the type of the agent is high enough, the platform retains her content even with the worst technology.

6.2.2 Regulator’s Sampling

I endogenize the regulator’s technology in the usual manner, defining $p$ as the ratio of the mass of violators to the mass of participants, i.e.,

$$p = \frac{\text{mass of violators}}{\text{mass of participants}}.$$

Proposition 6.2. A better platform technology results in a platform that retains a few high-type violators and both high and low-type compliants.

The proof is in the appendix.

Better technology reduces the platform’s error’s when evaluating both compliants and violators. In the case of compliants, improved technology lowers the threshold above which the user’s content is retained. This implies that as technology advances, the content of lower-type users is also retained. Conversely, if the technology becomes more adept at detecting violators, only the highest type violators are retained, as only their contribution to the platform’s revenue is substantial enough.

6.3 User liability

Consider a scenario where the regulator can identify and fine individual users, akin to regulations like NetzDG. In this setup, the platform can avoid penalties by reporting content that violates regulations, thereby shifting the burden of fines $f_u$ onto the individual users. However, the platform prefers to retain content rather than report it, as long as it can evade sanctions. This is because the platform generates revenue from user content. Consequently, top users anticipate that the platform will engage in cherry-picking, favoring their content in order to maximize revenue. To conclude, individual sanctions do not significantly alter the model’s outcomes. Note that since the model deals with a continuum of types, each type $x$ is drawn with 0 probability,

---

24 In a large, international context, such an assumption might seem unrealistic.
so a formal analysis is not possible. The intuition, however, is straightforward to highlight, as done above.

7 Additional application

The European Commission put forward a proposal concerning supply chain liability in 2022, titled The Directive on Corporate Sustainability Due Diligence [European Commission, 2022]. This directive institutes a corporate due diligence duty and introduces liability for European businesses regarding violations occurring within their supply chains outside the European Union.\footnote{https://ec.europa.eu/info/business-economy-euro/doing-business-eu/corporate-sustainability-due-diligence_en}

In the context of supply chain liability, consider a scenario involving a regulator, a firm, and its suppliers. As previously discussed, the firm’s profit increases when a supplier commits a violation, such as paying below the minimum wage to reduce production costs. The regulator inspects a supplier with an exogenous probability $p$, and, as in section 4.1, the firm’s highest profit is associated with the greatest societal harm.\footnote{Unlike in the platform scenario, the firm is unlikely to influence $p$} Mirroring earlier findings on cherry-picking, the firm optimally retains suppliers whose contribution to profit exceeds the potential sanction.

Improving the regulator’s approach to monitoring is straightforward: in anticipation of cherry-picking, the focus should be on inspecting suppliers that contribute most to the firm’s profits. Identifying these key suppliers could feasibly be achieved through the analysis of the firm’s financial accounting.

8 Conclusion

Online platforms play an indispensable role in today’s economic and social landscape. Yet, the increasing pressure on these platforms and their regulators to address illegal activities presents new challenges. This paper explores how regulators might enforce the liability of social media platforms for user-generated content. The theoretical model introduced here highlights ‘cherry-picking’, where platforms selectively penalize users to create an illusion of compliance and avoid sanctions. Cherry-picking becomes particularly detrimental when platforms favor users with vast audiences. While these users significantly boost the platform’s profits, their offenses can cause considerable societal harm.

To counteract cherry-picking, regulators could devise a mechanism linking the severity of sanctions to how platforms address offenses from users with smaller audiences versus those with...
larger followings. The optimal schedule for sanctions might not follow a straightforward pattern based on the offenses the regulator identifies on the platform. Specifically, the sanction may be higher when the regulator detects fewer violations on the platform. This result arises because, from the regulator’s perspective, there are two punishable actions: the offense committed by the user and the act of cherry-picking by the platform.

Another approach to curbing cherry-picking is for regulators to cease monitoring content deleted by the platform. However, this could lead to platforms deleting socially valuable content, either deliberately or accidentally, without any oversight or remedy from the regulator.

References


G. S. Becker and G. S. Stigler. Law Enforcement , Malfeasance , and Compensation of Enforcers

Author ( s ) : Gary S . Becker and George J . Stigler Source : The Journal of Legal Studies


L. Madio and M. Quinn. Content moderation and advertising in social media platforms â. (February), 2023.


C. Montag, B. Lachmann, M. Herrlich, and K. Zweig. Addictive features of social media/messenger platforms and freemium games against the background of psychological and economic theories, 7 2019. ISSN 1660-4601.


G. Sayantani. Twitter suspends several journalists, Musk cites ‘doxxing’ of his jet. pages 1–10, 2022.


A Proofs

A.1 Proof of Lemma 4.1

Proof. First, consider that the maximum fine \( \bar{f} \) is large enough, that is \( r(1, V) < \bar{f} \). Since \( \frac{r(x, V)}{x} > 0 \), a larger fine is required to incentivize the platform to delete content created by higher types. Hence, a fine that results in deleting the content of types \( x > x_s \) also yields deleting the content of type \( x < x_s \). Therefore the problem of the regulator has a corner solution, and it is either optimal to impose the maximum fine or no fine at all. Second, consider that the maximum fine is not enough to incentivize the platform to delete content of the top users. The above argument still holds why an interior solution cannot be optimal.

A.2 Proof of Lemma 4.2

Proof. The profit of the platform for type \( x_s \) is

\[
\pi(x_s, V, R, f_p) = r(x_s, V) - pf_p
\]  

(32)

Since the regulator’s objective is to deter violations for types \( x < x_s \), the optimal fine is

\[
f_p^* = \frac{r(x_s, V) + h_s(x_s)}{p}
\]  

(33)
Under $f_p^*$ the platform suspends only users with types lower than $x_s$. ■

A.3 Proof of Proposition 4.4

Proof. First, note that since the platform’s profit increases in the type of the user while the regulator does not condition the fine on the type of the user, the profit of the platform can be rewritten as

$$\pi(x,f_p) = \int_{R} r(x,V)g(x)dx + \int_{V} r(x,V)g(x)dx - \int_{V} f_p g(x)d(x)$$  \hspace{1cm} (34)

Where $R$ denotes the region where the firm suspends the user. In that region, the revenue is 0.

The platform deletes the content of the users at the bottom of the type distribution while allowing violations at the top. Denote with $x_p$ the largest type the platform deletes the content of. The platform’s problem becomes

$$\max_{x_p} \pi(x,f_p,x_p) = \int_{0}^{x_p} 0dx + \int_{x_p}^{1} r(x,V)g(x)dx - \int_{x_p}^{1} f_p g(x)d(x)$$  \hspace{1cm} (35)

The first order condition yields

$$\frac{\partial \pi(x,f_p,x_p)}{\partial x_p} = -r(x_p,V)g(x_p) + f_p g(x_p) = 0$$  \hspace{1cm} (36)

which can be rewritten as

$$r(x_p,V) = f_p.$$  \hspace{1cm} (37)

The second order condition is also satisfied,

$$\frac{\partial^2 \pi(x,f_p,x_p)}{\partial^2 x_p} = g'(x_p)(f_p - r(x_p,V)) - g(x_p) \frac{\partial r(x_p,V)}{\partial x_p} < 0.$$  \hspace{1cm} (38)

The interpretation of equation (37) is straightforward. The platform deletes the content of types whose contribution to the profit is less than 0, that is, for whom the fine is larger than the revenue from the content. That also means that an increase in fine decreases retention. Since the platform retains the content of the top users, cherry-picking is also increased with the larger fine.

$$x_p^* = r^{-1}(f_p,V)$$  \hspace{1cm} (39)

■
A.4 Proof of Proposition 4.5 and 4.6

Proof. The problem of the regulator writes as

$$\max_{f_p} W(x, f_p) = \int_0^{x_p(f_p)} 0dx + \int_{x_p(f_p)}^1 [r(x, V) + h_s(x)]g(x)dx$$  \hspace{1cm} (40)

The first order condition reads

$$\frac{\partial W(x, f_p)}{\partial f_p} = g(x_p^*) \left[ r(x_p^*, V) + h_p(x_s^*) \right] \frac{\partial x_p^*}{\partial f_p} = 0$$  \hspace{1cm} (41)

with $\frac{\partial x_p^*}{\partial f_p} = \frac{1}{r(r-1(f_p), V)}$. Hence to satisfy the first order condition for interior maximum,

$$r(x_p^*, V) = -h_s(x_p^*)$$  \hspace{1cm} (42)

Combining (42) and (37) yields

$$f_p^* = r(x_s^*, V)$$  \hspace{1cm} (43)

Note that the type for which the benefit and harms are equal is denoted by $x_s$. Hence it follows that the fine, $f_p$ is set such that $x_p^* = x_s$.

To check the sufficient condition the second derivative yields

$$\frac{\partial^2 W(x, f_p)}{\partial^2 f_p} = \frac{\partial g(x_p^*)}{\partial x_p^*} \frac{\partial x_p^*}{\partial f_p} \left[ r(x_p^*, V) + h_s(x_p^*) \right] \frac{\partial x_p^*}{\partial f_p} + g(x_p^*) \left[ \frac{\partial r(x_p^*, V)}{\partial x_p^*} + \frac{\partial h_s(x_p^*)}{\partial x_p^*} \right] \frac{\partial x_p^*}{\partial f_p} \frac{\partial x_p^*}{\partial f_p}$$

$$+ g(x_p^*) \left[ r(x_p^*, V) + h_s(x_p^*) \right] \frac{\partial^2 x_p^*}{\partial^2 f_p} \leq 0$$  \hspace{1cm} (44)

With $r(x_p^*, V) = h_s(x_p^*)$ the second order condition also becomes 0. However, inspecting the objective function yields a clear conclusion. If the violation is efficient for high types, $f_p = r(x_s, V)$ is the maximum of the objective. Low types for whom violation is inefficient are deterred, while high types are not. However, if violations of high types are inefficient, the above fine yields a minimum. Hence, the optimal fine has a corner solution, either allowing all users to violate or none of them. It depends on whether the platform generates social value if all users violate on the platform.
\[ W(x, f_p) = \int_0^1 r(x, V) + h_s(x)g(x)dx \geq 0 \quad (45) \]

\[ \square \]

A.5 Proof of Lemma 4.7

**Proof.** Given the information structure, the users can calculate the threshold on types, \( x_p(f_p) \) the platform applies for retaining content. Hence, they correctly anticipate that if their type is lower than \( r^{-1}(f_p, V) \) the firm would punish them. Therefore, in equilibrium, users self-select to be compliant if \( x < r^{-1}(f_p, V) \) and to be violators if \( x \geq r^{-1}(f_p, V) \). \[ \square \]

A.6 Proof of Lemma 4.9

**Proof.**

\[ \pi (x, f_p) = \int_V r(x, V)g(x)dx - n \int_V g(x)f_p d(x). \quad (46) \]

The first order condition reads

\[ r(x_p, V) = nf_p \quad (47) \]

It follows that the larger \( n \) is, the higher the choice of the platform \( x_p \) is. \[ \square \]

A.7 Proof of Lemma 5.1

**Proof.** Consider that \( f_2 \geq f_1 \). Take any choice of \( x_1 \) and \( x_2 \). By setting \( V'_p = V_1 + V_2 \), the platform’s profit is always larger under full cherry picking. Hence, it follows that the regulator chooses \( f_1 > f_2 \). \[ \square \]

A.8 Proof of Proposition 5.2

**Proof.** To simplify notation, I use \( r(x, V) = r(x) \). Take a threshold \( y \) such that the revenue of the platform is equal under full cherry picking and decreased cherry-picking, that is

\[ \int_0^{x_1} r(x)g(x)dx + \int_{x_2}^1 r(x)g(x)dx = \int_{x_2-y}^1 r(x)g(x)dx. \quad (48) \]

Hence \( y \) can be calculated from

\[ \int_0^{x_1} r(x)g(x)dx = \int_{x_2-y}^{x_2} r(x)g(x)dx \quad (49) \]

39
The right handside is decreasing in $y$. The larger $r(x_2)$ is, that is equivalent the steeper the function $r(x)$ is, the larger $y$ becomes. In other words, the more incentive the platform has to cherry pick, the smaller $x_2 - y$ is.

Intuitively, the larger the difference in the platform’s revenue between high and low types, the larger $y$ is, meaning a smaller increase in the violating area. Mathematically, since $r'(x) > 0$, $\Pr(x < x_2 - y) > \Pr(x_1 < x < x_2)$ that is

$$G(x_2 - y) > G(x_2) - G(x_1). \quad (50)$$

Moreover, $\Pr(x > x_2 - y) > \Pr(x > x_2)$, that is

$$G(x_2) > G(x_2 - y) \quad (51)$$

The probability of being fined with $f_1$ is $\Pr(f_1|\text{Fully cherry-picking}) = 2G(x_2 - y)(1 - G(x_2 - y))$ and $\Pr(f_1|\text{Decreased cherry-picking}) = 2(G(x_2) - G(x_1))(1 - G(x_2))$ under full-cherry picking and decreased cherry picking, respectively.

It follows from inequality (50) (51) that

$$\Pr(f_1|\text{Fully cherry-picking}) > \Pr(f_1|\text{Decreased cherry-picking}). \quad (52)$$

Similarly, $\Pr(f_3|\text{Fully cherry-picking}) = (1 - G(x_2 - y))^2$ and $\Pr(f_3|\text{Decreased cherry-picking}) = (G(x_1) + 1 - G(x_2))^2$. It follows form above that

$$\Pr(f_3|\text{Fully cherry-picking}) < \Pr(f_3|\text{Decreased cherry-picking}) \quad (53)$$

The platform may pay $f_2$ only under decreased cherry-picking.

Comparing the profit of the platform under the two strategies, (21) and (22) the platform prefers decreased-cherry picking if

$$(1 - G(x_2 - y))^2 f_3 + 2G(x_2 - y)(1 - G(x_2 - y)) f_1 \geq (G(x_1) + 1 - G(x_2))^2 f_3 + 2(G(x_2) - G(x_1))(1 - G(x_2)) f_1 - 2G(x_1)G(x_2) - G(x_1)) f_2 \quad (54)$$

Combining (54) with inequalities (52) and (53) yields that the regulator sets $f_1 > f_3$. ■
A.9 Proof of 5.3

Proof. The regulator’s problem can be written as

\[
\begin{align*}
\max_{f_1, f_2, f_3} W &= \int_0^{x_p^*} r(x)g(x)dx + \int_0^{x_p^*} h_s g(x)dx \\
\text{s.t.} & \quad \pi_1(x_p^*) \geq \pi_2(x_p^*) \\
& \quad f_1, f_2, f_3 \leq \bar{f}
\end{align*}
\]

Under \(\pi_1(x_p)\), the first order condition for the optimal cutoff point of the platform is

\[
\frac{\partial \pi_1(x_p)}{\partial x_p} = r(x_p)g(x_p) - 2f_3G(x_p)g(x_p) - 2f_2(1 - 2G(x_p))g(x_p) = 0
\]

and second order condition

\[
\frac{\partial^2 \pi_1(x_p)}{\partial x_p^2} = g'(x_p)[r(x_p) - 2f_3G(x_p)g(x_p) - 2f_2(1 - 2G(x_p))g(x_p)] + g(x_p)[r'(x_p) - 2g(x_p)f_3 + 2g(x_p)f_2] < 0
\]

Under \(\pi_2(x_p^*)\), the first order condition for the optimal cutoff point is

\[
\frac{\partial \pi_2(x_p^*)}{\partial x_p^*} = -r(x_p^*)g(x_p^*) - 2f_1(1 - 2G(x_p^*))g(x_p^*) + 2f_3(1 - G(x_p^*))g(x_p^*) = 0
\]

Increasing \(f_1\) decreases the optimal choice \(x_p^*\), while increasing \(f_3\) decreases it.

The second order condition is

\[
\frac{\partial^2 \pi_2(x_p^*)}{\partial x_p^*^2} = g'(x_p^*)[-r(x_p) - 2f_1(1 - 2G(x_p^*)) + 2f_3(1 - G(x_p^*))] + g(x_p^*)[-r'(x_p^*) + 2g(x_p)f_1 - 2g(x_p)f_2] < 0
\]

The Lagrangian of the regulator is

\[
\mathcal{L} = \int_0^{x_p^*} r(x)g(x)dx + \int_0^{x_p^*} h_s g(x)dx - \mu_0(\pi_2(x_p^*) - \pi_1(x_p^*)) - \mu_1 (f_1 - \bar{f}) - \mu_2 (f_2 - \bar{f}) - \mu_3 (f_3 - \bar{f})
\]

The first order conditions for an interior solution with \(i = \{1, 2, 3\}\)

\[
\frac{\partial \mathcal{L}}{\partial f_i} = g(x_p^*)[r(x_p^*) + h(x_p^*)] \frac{\partial x_p^*}{\partial f_i} - \mu_0 \left( \frac{\partial \pi_2(x_p^*)}{\partial x_p^*} \frac{\partial x_p^*}{\partial f_i} - \frac{\partial \pi_1(x_p^*)}{\partial x_p^*} \frac{\partial x_p^*}{\partial f_i} \right) - \mu_i = 0
\]
with complementary slackness conditions

\[ \mu_0(\pi_2(x_p^*) - \pi_1(x_p^*)) = 0 \quad (64) \]
\[ \mu_i(f_i - \bar{f}) = 0 \quad (65) \]

First, note that \( f_1 \) only decreases the profit under \( \pi_2 \) but does not decrease the objective function \( W \). Hence, by setting \( f_1 \) high the regulator can shift the incentives of the platform from cherry-picking to the higher social welfare option with profit denoted by \( \pi_1 \). \( f_2 \) only decreases the platform’s profit under \( \pi_1 \), therefore it is optimal to set \( f_2 = 0 \). Substituting \( x_s = x_p^* \) into (58) yields

\[ r(x_s) = 2f_3G(x_s) \quad (66) \]

or

\[ f_3 = \frac{r(x_s)}{2G(x_s)} \quad (67) \]

Which is the first best value of \( f_3 \). However, this may not be achievable. Proposition 5.2 establishes that \( f_1 > f_3 \) and

\[ \text{Pr}(f_3|\text{Fully cherry-picking}) < \text{Pr}(f_3|\text{Decreased cherry-picking}) \quad (68) \]

Since \( f_3 \) is constrained by \( f_1 \), the regulator may not be able to set \( f_3 \) high enough since that would countereffect the impact of \( f_1 \) on cherry-picking. From (58) it follows that a lowering \( f_3 \) yields a higher threshold, \( x_p^* \) and hence under deterrence occurs. Overdeterrence is not possible, since if \( f_3 \) is too high, the regulator would decrease it to the first best level, which must also satisfy (56).

A.10 Proof of 6.2

Proof. The platform’s expected profit when all agents violate on the platform is

\[ \pi(x, f_p, q_V) = \int_V q_V r(x, V) g(x) dx + \int_V (1 - q_V) r(x, C) g(x) dx - \int_{V_p} q_V g(x) f_p d(x) \quad (69) \]

As established in the previous section, the platform follows a threshold strategy, that is the platform retains the content only if the agent’s type is high enough. Hence, the objective reads
similar to (35).

\[
\pi(x, f_p, q_V) = \int_{x_{p(q_V)}}^{1} q_V r(x, V) g(x) dx + \int_{x_{p(q_V)}}^{1} (1 - q_V) r(x, C) g(x) dx - \int_{x_{p(q_V)}}^{1} q_V g(x) f_p d(x)
\]

(70)

The first order condition writes

\[-q_V r(x_p, V) - (1 - q_V) r(x_p, C) + q_V f_p = 0\]

(71)

Since \( r(x, V) > r(x, C) \), the better the technology of the firm, represented by \( q_V \), the higher the threshold, \( x^*_p(q_V) \) becomes. The better technology means the firm is more certain than agent is violating therefore only contents of high types are retained.

Next, consider that all agents are compliant on the platform.

\[
\pi(x, f_p, q_C) = \int_{C} q_C r(x, C) g(x) dx + \int_{C} (1 - q_C) r(x, V) g(x) dx - \int_{C} (1 - q_C) g(x) f_p d(x)
\]

(72)

As established before, the firm follows a threshold strategy with compliants too. If the the type of the agent is high enough, the platform retains her content.

\[
\pi(x, f_p, q_C) = \int_{x_{p(q_C)}}^{1} q_C r(x, C) g(x) dx + \int_{x_{p(q_C)}}^{1} (1 - q_C) r(x, V) g(x) dx - \int_{x_{p(q_C)}}^{1} (1 - q_C) g(x) f_p d(x)
\]

(73)

With first order condition

\[-q_C r(x_p, C) - (1 - q_C) r(x_p, V) + (1 - q_C) f_p = 0\]

(74)

The threshold \( x^*_p(q_C) \) becomes smaller the better the technology is. That is, the more certain the platform is that the users are complying with the rules, the more content the platform retains.

Finally, consider that with \( z \) probability all agents are violent, and with \( 1 - z \) probability all agents are compliant. \(^{27}\) The platform’s objective is then

\(^{27}\)Similar is if \( z \) proportion of the agents are violent.
\[
\pi(x, f_p, q_V, q_C, z) = z \left[ \int_{x_{pqv}}^{1} q_V r(x, V) g(x) dx + \int_{x_{pqv}}^{1} (1 - q_V) r(x, C) g(x) dx - \int_{x_{pqv}}^{1} q_V g(x) f_p d(x) \right] \\
(1 - z) \left[ \int_{x_{pqc}}^{1} q_C r(x, C) g(x) dx + \int_{x_{pqc}}^{1} (1 - q_C) r(x, V) g(x) dx - \int_{x_{pqc}}^{1} (1 - q_C) g(x) f_p d(x) \right].
\]

(75)

The objective yields identical first order conditions to (71) and (74), hence leads to the same result as previously.