Efficient copyright filters for online hosting platforms

Alessandro De Chiara[*], Ester Manna[†], Antoni Rubí-Puig[‡], Adrian Segura[§]

April 23, 2024

Abstract

We build a model where an online hosting platform develops a copyright filter to screen content that contributors wish to upload. The technology is imprecise, since non-infringing material may be incorrectly filtered out. Once the content is hosted on the platform, a right-holder may send a take-down notice if its own monitoring system, also imprecise, finds it to be copyright infringing. The efficient design of regulation and liability calls for (i) giving the right-holder incentives to evaluate fair use when submitting a notice and (ii) lifting the safe-harbor protection granted to platforms that promptly remove content following a take-down notice.

Keywords: Article 17; Copyright filters; Fair use; Infringing material; Liability rules; Notice and take-down system; Online hosting platforms.

JEL classifications: K2; L51.

[*] Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: aledechiara@ub.edu.

[†] Professora Lectora Serra Húnter, Department of Economics, Universitat de Barcelona and Barcelona Economic Analysis Team (BEAT), Avinguda Diagonal 696, 08034, Barcelona, Spain. E-mail: ester-manna@ub.edu.

[‡] Department of Law, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27 08005 Barcelona. E-mail: antoni.rubi-puig@upf.edu.

[§] Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas, 25-27 08005 Barcelona. E-mail: adrian.segura@upf.edu.

1 Introduction

It is a well-known fact that copyright enforcement has become more challenging in the digital age. More than twenty years ago, the controversial Section 512 of the *Digital Millennium Copyright Act* (DMCA) ushered in the notice and take-down process in a bid to foster cooperation between online hosting platforms (OHPs) and right-holders. In a nutshell, OHPs can escape liability if they promptly remove copyright-infringing works from their websites following a take-down notice sent by the copyright holders. These, in turn, need not go through all the hassle and expenses associated with a lawsuit.

Over the past decades, OHPs have risen to prominence, gaining both power and influence. As a result, new regulatory rules have been proposed and adopted on both sides of the Atlantic in an effort to catch up with the radical changes in the digital landscape. Most notably, the Article 17 of the *2019 EU Directive on Copyright in the Digital Single Market* establishes the rights and the obligations of both the OHPs and the right-holders for the use of copyright-protected works. In particular, while excluding general monitoring obligations, OHPs can avoid liability if they can demonstrate they made the best effort to ensure unavailability of protected works. The common thrust of these new rules is that of making OHPs responsible for the content that they host, as also highlighted by the recently approved 2020 Digital Services Act.[1]

Although the objective of these provisions may be meritorious, so far the implementation of rules meant to improve copyright enforcement has not gone without a hitch. In particular, there is a large debate over the negative consequences associated with the adoption of the notice and take-down system. Urban et al. (2017a,b) extensively document and discuss the excessive number of flawed notices typically sent by large right-holders' own automated system coupled with the OHPs' high cost of assessing the accuracy of a received notice and the limited incentives to challenge it. Moreover, although there is no general monitoring obligation in place, some large OHPs have adopted ex-ante filter systems to screen content before it is made available online and this may also lead to type-I errors.[2] Arguably, the most notable examples of such copyright filters are Youtube's Content ID and Audible Magic's own automated content recognition technology.

It is not clear that the currently discussed changes will mend the aforementioned side-effects. In this paper, we build an economic model that accounts for the issues inherent in the adoption of copyright filters and the working of the notice and take-down system to investigate the efficient design of regulation and liability. In our model, an OHP can

[1] For a description of the current EU liability regime for online hosting platforms, its issues, and the policy proposals that are being discussed see Madiega (2020).

[2] E.g., see Frosio (2017) who reports and elaborates on the position of the European Court of Justice that automated filters cannot replace human judgment.

develop a filter to detect whether the content that contributors want to upload infringes material protected by copyright. The technology is imperfect in that the filter may mistakenly block content that is not actually infringing from being uploaded. Through this assumption, we mean to capture one critical feature of the automated filter technologies adopted by OHPs. Namely, their possible failure in distinguishing between copyright-infringing content and content that makes fair use of existing material or, otherwise, is covered by an exception or limitation to copyright.[3] The magnitude of this issue is more severe the stricter the filter developed by the OHP. Once the content is hosted on the platform, a right-holder may send a notice if its own automated notice system finds it to be copyright infringing. We posit that this automated system is imperfect too as the content identified as infringing may in fact represent fair use. The OHP can either accept or challenge a received notice. In the former case, the content will be made unavailable to users whereas, in the latter case, a third party will adjudicate the dispute.

Regulation and liability rules should aim at minimizing the cost of achieving an efficient copyright enforcement and safeguarding contributors who make fair use of existing material. As we show, this requires that: (i) the OHP take a proactive role in filtering out the copyright infringing material without overly excluding contributors who make fair use; (ii) the right-holders do not send an excessive number of inaccurate notices. We find that these two objectives are closely intertwined and solely imposing liability on OHPs for copyright infringement, without punishing right-holders who send inaccurate notices, may backfire. More specifically, introducing a penalty in the case in which a court upholds the OHP's decision to challenge a notice can be desirable. Such a penalty can induce right-holders to seriously evaluate whether the hosted material makes fair use before submitting a notice. The size of this penalty must be positively related to the damages that a right-holder would recover should the court instead agree with its claim. This stands in stark contrast with the current practice of imposing substantial penalties

[3] For the purposes of this article, no distinctions are made between jurisdictions that establish legal exceptions and limitations to copyright (e.g., the EU copyright law) and jurisdictions that, in addition, resort to a general fair use defense (e.g., the US federal copyright law). Even though there is an extensive literature exploring the differences between rules and standards, that points out to differences in predictability and legal certainty, this distinction is less acute in the field of copyright law. First, the usual situations - such as parodies, caricatures, quotations or criticism -, are covered in both systems either through a statutory exception established by the legislator or through case-law interpreting the fair use standard. Second, concepts included in statutory exception rules can also be affected by ambiguities and uncertainty and require interpretation and flexibility. And finally, some scholars have argued that courts reduce uncertainty by elaborating the fair use standard into crystalized rules (see Elkin-Koren and Fischman-Afori, 2017). In the end, automated filter technologies will have to detect whether a particular content is, for instance, a lawful quotation irrespective of whether quotations are covered by an exception or limitation or are considered fair use. In the remainder of the article, we use the expression *"fair use"* to refer to both situations.

on the OHP if the court finds it to host infringing works, while essentially making it unlikely for the OHP to recover any damages from the right-holders. In fact, the envisioned system would not trigger excessive litigation, thereby avoiding its associated costs. This is because the OHP would have an incentive to challenge notices that are deemed to be inaccurate and, consequently, the right-holders would exert care when submitting notices. Exactly because there are not too many inaccurate notices, they can be used along with the actual number of take-down decisions as a signal for the regulator to evaluate the accuracy of the copyright filter adopted by the OHP. More in detail, the regulator can condition its intervention on the take-down to notice ratio or on the ratio of content taken down following right-holders' notices over the total content hosted in the OHP. The OHP should incur some penalty when this ratio is above some pre-specified threshold. This measure would essentially imply that no safe harbor protection would be granted to well-established OHPs that promptly take down content following a notice, as they would also need to take on a pro-active role in filtering out uploaded content. In conclusion, our proposed solution involves both regulation, that works ex-ante, and courts (or other mechanisms for the adjudication of disputes), that could be called to intervene ex-post. This dual system jointly achieves efficient copyright enforcement without excluding fair-use material. For this solution to properly work, the regulator must acquire a great deal of information. Part of it will have to be directly provided by the parties involved (i.e., the OHP and the copyright holder), whereas other should be obtained by the regulator through meticulous and independent analyses, aiming for instance at ascertaining the right-holder's harm due to an undetected infringement.

Outline. The rest of the paper proceeds as follows. In the next subsection, we review the related literature. In Section 2, we develop the baseline set-up that is analyzed and solved in Section 3. In Section 4, we extend the analysis of the model by allowing the right-holder's own automated system to be imprecise. In Section 5, we consider some robustness checks and extensions. In Section 6, we discuss future directions for our research and provide concluding remarks and policy implications.

1.1 Related Literature

Our paper relates to the economics literature that considers platforms as intermediaries that facilitate the interaction between different user groups and study their regulation.[4] Recent papers have analyzed questions of platform governance, that is, the platform's design of rules that govern the relationship between the different platform user groups

[4] See Jullien and Sand-Zantman (2021) for a definition of platforms and a discussion of topical competition policy issues.

(prominent examples include Johnson et al., 2020, Johnen and Somogyi, 2021, and Teh, 2021). This is an issue that has also attracted significant policy interest (e.g., see Section 4.III of Crémer et al., 2019). The platform's choice of the filter technology that affects which content is made available online is indeed one relevant example of platform governance. Our work is especially related to Casner (2020) who considers a platform that can screen sellers depending on their quality, highlighting a strategic incentive to keep average seller's quality low. We study how the platform's screening effort is shaped by regulation and users' valuation of content originality (see Section 5.1). The way online content is moderated is a hotly debated topic nowadays: Jiménez Durán (2021) shows through an experimental setup that randomly reported posts on Twitter for violating conduct speech rules increases the probability that Twitter deletes the disputed content; Zhang (2021) finds that enforcing take-down policies on GitHub results in efficient welfare allocations; lastly, Madio and Quinn (2023) find that a platform's incentives to moderate content are higher when users and advertisers have congruent preferences.

There is surprisingly little in the economics literature concerning the efficient design of regulation and liability of platform's copyright filtering technology and notice and take-down system, despite their great relevance in the online sharing economy. A notable exception is represented by Buiten et al. (2020)'s paper in which the authors analyze the platforms' incentives to host infringing material and describe the efficient liability rule for hosting services in the European Union. Recently, Lefouili and Madio (2022) describe how a stricter liability rule affects several key variables, such as online platforms' prices and investments. However, both papers do not develop a formal model to account for these issues. In the management literature, Jain et al. (2020) study the monitoring incentives of a right-holder and an online platform when users can decide to consume illegal content. We offer a different perspective to a similar topic by focusing on the design of regulation and liability. More closely related to our paper is Jeon et al. (2021) who study a platform's incentive to screen out trademark-infringing content and its effect on sellers' innovation investments. The modeling set-ups are sharply different as the two papers aim to capture features of different types of platforms: their model is better suited to study online marketplaces and app stores, whereas our paper better applies to online hosting platforms, such as YouTube. More central to our paper is the issue associated with false positives when screening out infringing content, whereas competition in the online marketplace between brand owners and low-quality IP-infringing sellers has a more prominent role in their analysis. There is also a clear link with the recent work by Hua and Spier (2021) who study a two-sided platform that is an essential intermediary between potentially harmful firms and users. The platform can screen firms through a costly audit technology and the interaction price. The authors find that platform liability is critical to providing socially optimal auditing incentives when firms are judgment proof. In our set-up, potentially harmful contributors have limited assets and the right-holder's harm is not always verifiable. The right-holder plays an instrumental role in detecting harmful content and we study how to induce the platform to make the socially efficient investment in its filter technology. We also find that right-holder's liability is necessary to improve the accuracy of the take-down notice requests.[5]

Our paper touches on the issue of indirect liability for copyright infringement, which has been studied by Landes and Lichtman (2003). They argue that the rationale to impose liability onto the platforms is that they can monitor the level of care exercised by contributors and make it cheaper for copyright holders to sue platforms rather than sue multiple potential infringers. How the provision of safe harbors affects market structure is studied by Beard et al. (2018). In their model limited liability for online platforms leads to a unique equilibrium in which the most profitable platforms are those with high levels of illegal material. A separating equilibrium may arise if the risk of liability is increased. Under this scenario, platforms offering online legitimate content will coexist with the ones offering only low quality or infringing material.[6]

The legal literature has paid considerable more attention to the problems associated with copyright enforcement in the digital age. In particular, the controversial Section 512 of the DMCA and the subsequent directives were followed by several legal studies that have discussed the role and responsibilities of online hosting platforms. Urban et al. (2017a) have been quite influential in showing the costs that the current regime imposes on platforms and contributors as well as its inefficiencies. Gabison and Buiten (2019) argue that imposing liability on online platforms would induce them to internalize the costs of copyright infringements. Hornik and Villa llera (2017) address the efficiency of liability rules under the 2000 E-Commerce Directive. They point out that a strict liability regime may restrict freedom of expression. Nevertheless, a no liability regime would transfer all the negative consequences of infringing content from hosting platforms towards the society.[7] Grimmelmann and Zhang (2023) introduce a framework for understanding intermediary liability. Their model emphasizes key elements such as externalities, imperfect

[5] Other recent papers on platform liability include: Hua and Spier (2023) who compare the efficiency of strict liability and negligence in the presence of network externalities; Zennyo (2023) who shows that a platform has no incentives to voluntarily adopt liability for third-party harm; De Chiara et al. (2023) who study the interplay between platform liability and reputational sanctions; and Feher (2023) who highlights a platform's incentives to only punish content creators that cause little harm and bring small profits.

[6] Other relevant papers in the law and economics literature that have studied OHP's liability and safe harbor are Cotter (2006) and Liebowitz (2018).

[7] See Schruers (2002) for an analysis of the role of liability rules and their evolution in the US context from a law and economics perspective. He concludes that the main solutions have traditionally failed to provide an accurate level of monitoring.

information, and investigation expenses. When a platform hosts user-generated content, it faces the challenge of discerning which content might be beneficial. The platform is uncertain about whether the content is harmful or not but can assess the probability of harm associated with each piece of content. Based on this probability, the platform chooses to remove or not the content. The results mimic the inherent tradeoffs in content moderation: false positives v. false negatives and costly but accurate v. less costly but less accurate reviewing processes. We complement these studies with a formal economic analysis of the interplay between different actors that are affected by copyright enforcement in online platforms.

2 Set-up

We consider a model in which *users* can join a *platform* to enjoy some content. A *right-holder* may suffer some harm depending on the type of content posted online.

Platform. The platform can host content that is provided by some external (non-modeled) contributor. It is common knowledge that the developed content is original ($\theta = 1$) with probability $\beta \in (0, 1)$ and unoriginal ($\theta = 0$) with complementary probability $1 - \beta$. The platform cannot costlessly observe the originality of the content, but can invest in a filter that excludes unoriginal content. This investment is denoted $e \in [0, 1]$ and costs $\psi(e)$, which is increasing and strictly convex. If there is no filter, i.e., $e = 0$, every content is admitted. This means that without a filter all items that are indeed original are approved (no type-I errors), but all items that should be rejected because they are unoriginal are also admitted (there are type-II errors). The use of a filter reduces the occurrence of type-II errors but also leads to some type-I errors. In particular, we model the filter technology as follows: the filter generates a binary signal on the item submitted by the contributor, $s \in \{0, 1\}$, where $s = 0$ means that the content is unoriginal and should be rejected. Assume the following conditional probabilities: (i) $Pr[s = 0|\theta = 0] = e$ and (ii) $Pr[s = 1|\theta = 1] = 1 - \gamma e$, with $\gamma \in (0, 1]$. Because of (i), a higher investment in the filter e increases the probability that unoriginal content is rejected; at the same time, due to (ii), a higher investment in the filter may lead to some original content being wrongfully rejected. The magnitude of this side effect positively depends on γ. We henceforth say that the filter is stricter when e takes a higher value.[8] The platform enjoys some

[8] It is important to stress that our modeling assumption on the positive relationship between strictness of the filter and Type-I errors is in line with the experts' views on the accuracy issues inherent in digital fingerprinting techniques and the tension between false positives and false negatives (e.g., see the 2020 Report by the European Union Intellectual Property Office "Automated Content Recognition: Discussion Paper – Phase 1 'Existing technologies and their impact on IP").

advertising revenues that are increasing in the mass of users who join the platform, where the latter is denoted D. The platform's profit is:

$$\pi = a\ D\ \mathbb{1}_A - \psi(e),$$

where a is the per-unit advertising revenue, and $\mathbb{1}_A$ is an indicator function that takes value 1 if the content is approved and posted online, and 0 otherwise.

Users. There is a unit mass of users who derive benefits from enjoying the content on the platform. In particular, the utility of the generic user i is:[9]

$$V_i = v_i\ \mathbb{1}_A,$$

where v_i represents user i' utility from the content. This may be negative to capture the opportunity cost of spending leisure time in that particular way instead of some other activity (e.g., reading a book, going for a walk). Users are heterogeneous in this opportunity cost. Specifically, v_i is distributed according to the commonly-known continuous distribution function F on the interval $[\underline{v}, \overline{v}]$, with $\underline{v} < 0 < \overline{v}$.

Right-holder. A copyright holder suffers some harm valued $H > 0$ if the unoriginal content is kept online. Unoriginal content that is made available on the platform is detected with probability $\eta \leq 1$, in which case the right-holder sends a notice to the platform. Upon receiving a notice, the platform bears a processing cost $k \in (0, H)$. In practice, some human resources may have to be assigned to evaluate the right-holder's request (e.g., to check whether the take-down notice has been correctly filed, to assess the strength of the right-holder's claim, to inform the contributor). This cost may also include the related capital expenses, such as the technology needed to review the notice. If content is removed, users do not enjoy utility and the platform does not obtain advertising revenues. In the baseline model, the right-holder is a passive player that does not take any action. We relax this assumption in Section 4, in which we bring attention to the copyright holder's problem of checking ex-post whether some unoriginal content has been posted. In doing so, we will also allow for right-holder's type-I errors and we study their interplay with the platform's choice of the filter technology.

Timing of the game. The sequence of events is as follows:

1. The platform invests $e \in [0, 1]$ in the filter technology.

[9] For now, users' utility is directly unaffected by the originality of the content. We relax this and other assumptions of the baseline model in Section 5.

2. The platform observes a signal on the originality of the content. If $s = 1$, the content is allowed on the platform; if $s = 0$, the content is rejected and the game ends.

3. Users decide whether or not to join the platform, without knowing whether the content is indeed original or not.

4. With probability η, unoriginal content is detected and the platform can remove this content. All players derive utility.

The equilibrium concept we employ is Perfect Bayesian equilibrium. All players form beliefs about the type of the other players they face. All proofs are in appendix.

Objective. We aim to study how a legislator who maximizes social welfare would design the liability regime. We assume that the legislator can impose payments onto the parties and can set damages awards.

Discussion. For simplicity, in the model we use the dichotomy original versus non-original. To avoid ambiguity, it is worth clarifying what we mean with these two terms. By original content we broadly refer to material that represents an intellectual creation of the contributor and may be either some novel work or some creative reinterpretation or adaptation of existing work. In the latter case, it would thus constitute *fair use* of existing material. By non-original content we refer to content that is neither novel nor represent a marked departure from existing work.

We have implicitly assumed that advertising revenues are large enough so that the platform is unwilling to set a positive price that users have to pay to access the content.[10] We will discuss the robustness of our conclusions when the platform sets a positive per-user fee in Section 5.5.

To keep the model simple and focus on the platform's and right-holder's incentives to screen content to determine their originality, we have refrained from modeling the contributor. We think of contributors as users who may derive some benefit from having their material posted online. Thus, the socially-minded legislator would take into account their utility when designing the liability rule. Specifically, we will assume that there

[10] For a monopolistic platform, this would require that $a \geq \frac{1-F(0)}{f(0)}$ if $\frac{1-F(\cdot)}{f(\cdot)}$ is decreasing in its argument. To see why, consider a more general profit function for the platform in Stage 2: $\pi = [aD + pD]\mathbb{1}_A$, where p is the per-user fee and $D = 1 - F(p)$. The first-order condition yields:

$$p = \frac{1 - F(p)}{f(p)} - a,$$

which is the standard monopoly pricing condition, if it were not for the term a. If a is large enough, the platform would set $p = 0$. In fact, it would even subsidize users' entry by setting a negative price. We abstract away from this possibility.

is some inalienable benefit $B \geq 0$ that is derived by these contributors whenever their content is approved and kept online. This is meant to capture the ego boost and monetary benefits (e.g., corporate sponsorship, merchandise sales) that a contributor may obtain from sharing content online. We allow for cross-group externalities in Section 3.3, since such benefit may well depend on the number of users who join the platform - and the users' benefits may be positively related to the amount of posted content. Another, related question concerns the incentives that may be provided to contributors in order to motivate content development. We defer a discussion of this and other issues to the conclusions. Throughout, we assume that contributors are both liquidity constrained and judgement proof. As a result, contributors cannot be held liable for the harm they may cause to the right-holder and cannot be charged a price to upload content to the platform. This indeed provides a strong (non-formal) argument for making a platform liable for the content it hosts (see also Hua and Spier, 2021).

To provide a real-world example where our model could apply, we have in mind a social-media or an online video-sharing platform that follows an ad-based business model. In such a platform, some users can decide to upload content, thereby becoming what we call contributors, but this material may turn out not to be original, infringing copyright.

Benchmarks. We now study two benchmarks. First, if there is *no liability*, it is straightforward to show that the platform would have no incentive whatsoever to remove unoriginal content after receiving a notice or to invest in a filter technology. This is because a stricter filter would just decrease the amount of content available online, thereby reducing the platform's profit.

Consider now *first-best*. Let users' demand be:

$$D := \int_0^{\bar{v}} f(x)dx,$$

and let the overall utility of the users who join the platform be:

$$V := \int_0^{\bar{v}} F(x)dx.$$

In the absence of asymmetric information, unoriginal content should be prohibited if $H > aD + V + B$, that is, if the disutility suffered by the right-holder outweighs the benefits accruing to the platform, the users, and the contributor when the content is kept online.[11] For now, we assume that $H > aD + V + B + k$ so that costly content removal is always optimal when such content would harm the right-holder. We relax this assumption in Section 5.4.

[11] In Appendix B, we show that a legislator that maximizes social welfare may be willing to commit to copyright protection even when this inequality does not hold.

3 Efficient Copyright Filter

We start this section by pinning down the filter technology that would be chosen by a social planner to maximize welfare in the presence of information asymmetries. We then study how liability rules should be designed to reach or at least approximate this second-best efficient solution. We also caution on the challenges that a practical implementation of a liability rule may encounter to achieve efficiency and we propose some potential remedies. Lastly, we discuss different objectives that a legislator may pursue in designing the liability rule and how to achieve them.

3.1 Second best

We now determine the legislator's choice of the filter technology in the presence of asymmetric information (second best, that we henceforth denote SB). Advertising revenues enter the welfare expression because we assume that they are not a mere transfer among players but advertising leads to a better match between consumers and products. We abstract away from nuisance costs imposed by advertising on consumers.

Before illustrating the social welfare function, we note that the probability that content is posted online depends on the strictness of the filter technology e and is given by:

$$\mathbb{1}_{A(e)} := Pr[A] = \beta(1-\gamma e) + (1-\beta)(1-e).$$

Original contributions are approved with probability $1-\gamma e$, whereas unoriginal ones are approved with probability $1-e$. In second best, the legislator would choose $e \in [0,1]$ to maximize the following social welfare expression:

$$\begin{aligned} W^{SB} = & \overbrace{[aD+V+B][\mathbb{1}_{A(e)} - (1-\beta)(1-e)\eta]}^{\text{Benefits of accepting and keeping the content online}} \\ & - \underbrace{[(1-\beta)(1-e)\eta k + (1-\beta)(1-e)(1-\eta)H]}_{\text{Cost of posting the copyright-infringing content}} - \underbrace{\psi(e)}_{\text{Filter cost}} . \end{aligned} \tag{1}$$

The first term in square brackets is the sum of platform's (aD), users' (V), and contributor's (B) utility if the content is approved. The probability that the content is approved is $\mathbb{1}_{A(e)}$, that is, the likelihood that the content successfully passes through the filter, minus $(1-\beta)(1-e)\eta$, which is the probability that an infringement is ex-post detected in which case the content is optimally removed.[12] The terms on the second line represent, respectively, the cost of posting a copyright-infringing content online (event that

[12] Throughout the paper, we will assume that the time window between Stage 3 and Stage 4 is so short that users would not derive utility from enjoying the content and, accordingly, the platform would not obtain advertising revenues. It is worth remarking that our results would not be qualitatively altered if part of users' and firms' benefits were obtained between the two stages.

has probability $(1-\beta)(1-e)$), which is detected with probability η and goes undetected with probability $1-\eta$, giving rise to a loss equal to k and H, respectively. The very last term is the cost of developing the filter technology. Let e^{SB} be the second-best level of the filter technology. At an interior optimum, this is derived implicitly from the following first-order condition:[13]

$$\begin{aligned}&-[aD+V+B][\gamma\beta+(1-\beta)(1-\eta)]\\&+(1-\beta)\eta k+(1-\beta)(1-\eta)H=\psi'(e^{SB}).\end{aligned} \tag{2}$$

The terms on the first line represent the downside of marginally tightening the filter for the platform, the users, and the contributor: the probability of having some content online decreases. The upside of having a marginally stricter filter is instead captured by the first two terms on the second line: a lower probability that unoriginal content that harms the right-holder is posted online, and either goes undetected or is costly removed. The term on the right-hand side is the direct cost of setting a marginally stricter filter. Higher values of H and k call for a stricter filter, whereas a higher ability of the right-holder to catch unoriginal content ex-post, i.e., a higher value of η, has a negative effect on the efficient strictness of the filter.

3.2 Achieving second best

The aim of this subsection is to study how the legislator could achieve second-best and to highlight some challenges that it may practically face.

A thorny question concerns the platform's incentive to develop the socially desirable filter technology. The envisioned system includes a notice and takedown process, which allows the platform to eschew liability by removing content that a copyright holder claims to be infringing. As long as the cost of removing such content is lower than the expected cost of liability, the platform will always prefer to take the content down. Therefore, unless the damages the platform may be forced to pay are relatively small, on-the-equilibrium path the platform will never have to pay damages as it would never challenge a notice. As a consequence, such liability rule may have a suboptimal effect on the platform's choice of the filtering technology. To show this, suppose that the platform has to pay damages H to the copyright holder if an infringement is detected and is not taken down, whereas it does not face damages otherwise. As $k<H$, content is always taken down following a notice.[14] The platform's maximization problem is:

$$\max_{e\in[0,1]} \quad aD[\mathbb{1}_{A(e)}-(1-\beta)(1-e)\eta]-(1-\beta)(1-e)\eta k-\psi(e).$$

[13] For $e^{SB}>0$, it must be that k and H are sufficiently large as compared to aD, V, and B.

[14] Recall that we are assuming that the notice always follows a detected infringement and this ex-post infringement technology does not entail any type-I errors. We will examine the implications of relaxing this assumption in the next section.

At an interior optimum, the first-order necessary condition yields:

$$-aD[\gamma\beta + (1-\beta)(1-\eta)] + (1-\beta)\eta k = \psi'(e^{**}). \tag{3}$$

By comparing (3) with (2), it is possible to see that the filter technology chosen by the platform in the presence of this negligence rule will generally not be second-best efficient. This is because the platform does not internalize the benefits that the contributor and the users obtain when the content is available online, nor the harm the right-holder suffers when infringing content is posted online and goes undetected. The wedge between the platform-chosen filter and the optimal one will be larger the greater the value of H and the smaller η. This is because better ex-post detection unequivocally leads to a higher platform's choice of e. Below, we suggest a way to induce the platform to choose the socially desirable filter technology.

Liability rule implementing second best. The gist of this solution lies in removing the provision that grants safe harbor protection to a platform that has promptly removed content following a notice. As some of the illegal content remains undetected (i.e., as long as $\eta < 1$), the copyright holder suffers some harm for which it is not compensated. The legislator could use the information conveyed by the accepted take-down notices to induce the platform to choose $e = e^{SB}$. This could be achieved via a *strict liability rule*. For the sake of simplicity, in the model we now assume that, whenever some copyright infringement is detected, it results in an expected loss valued d to the platform: this is given by the probability that the legislator intervenes multiplied by the requested payment, which is made directly to the legislator and is therefore welfare neutral. The value of d is endogenously determined in the model. The platform would now solve the following objective function:

$$\max_{e\in[0,1]} \quad aD[\mathbb{1}_{A(e)} - (1-\beta)(1-e)\eta] - (1-\beta)(1-e)\eta(k+d) - \psi(e).$$

The following proposition illustrates how to set d so as to achieve second best.

Proposition 1. *To achieve second-best efficiency, the legislator should set d to make the platform internalize the harm that a copyright holder suffers when the infringement goes undetected, as well as users' utility and the contributor's private benefit:*

$$d^* = \frac{(1-\eta)H}{\eta} - \frac{(V+B)[\gamma\beta + (1-\beta)(1-\eta)]}{(1-\beta)\eta}. \tag{4}$$

In principle, the optimal expected payment to the state d^* could be negative. This is more likely to be the case when users' and contributor's benefits, V and B, respectively, are larger. It follows that the platform's ability to extract V and B would increase d^*.[15] On the other hand, d^* could be greater than H: a necessary condition for this to occur is that $\eta < 1/2$, that is, the technology used to detect unoriginal content available on the platform must be somewhat imprecise. More in general, when ex-post detection is less accurate (i.e., η takes smaller values), the expected payment d will have to increase to align the platform's screening incentives with the legislator's if $H > (B+V)\left[1+\frac{\beta\gamma}{1-\beta}\right]$. When the content is more likely to be original (i.e., β is higher) or the filter technology makes more frequently type-I errors (i.e., γ is higher), then the optimal payment decreases. Intuitively, having a looser filter is preferable when original content is often incorrectly excluded by the platform's filtering algorithm.

3.3 Discussion

Below, we discuss some of the features and implications of the policy outlined in Proposition 1, as well as its feasibility.

Punitive damages. Imposing a penalty on the platform to achieve second best may resemble the idea of punitive damages to induce a tortfeasor to take optimal precautions advocated by Polinsky and Shavell (1998), but it is different in several aspects. Firstly, in our model, the payment d could also be negative as the mechanism may also work for overdeterrence. Secondly, and more importantly, the payment is imposed to a third-party, which is not necessarily the direct tortfeasor in the relationship, and its size depends on the platform's ability to detect infringements. Conversely, in Polinsky and Shavell (1998) the economic rationale behind punitive damages is to provide firms with incentives to be compliant.

Implementation. There are a plethora of ways to implement the second-best in practice. The legislator could intervene and impose a small payment onto the platform whenever a notice holds up. Alternatively, the legislator's intervention could be triggered if the fraction of accepted notices over the overall content that is made available on the platform exceeds some acceptable, predetermined threshold. After all, a relatively large number of take-downs would suggest that the platform's filter technology does a poor job in preventing copyright infringing content from being posted. The higher the selected threshold, the larger the payment that the legislator should impose on the platform if the threshold is exceeded. The solution would thus impose a series of procedural obligations

[15] It is worth remarking that, if the platform could charge a price to a contributor for uploading material, it could potentially exclude unoriginal content. In this regard, see Hua and Spier (2021) who show under what conditions the price charged by the platform helps screening out harmful firms.

onto online platforms, which would be required to disclose detailed information about the number of notifications received, as well as the way they have handled them. In particular, whether the notification eventually led to content being removed or was successfully challenged.[16,17]

Rather than introducing the payment d, the legislator could increase the platform's cost of processing take-down notices k. For instance, this can be achieved by mandating strict transparency rules on notice decisions, in a fashion similar to what the Digital Services Act proposes for platform's content moderation decisions. Even though raising the cost of handling notices ex-post would prompt the platform to dedicate more resources to ex-ante screening, this approach would not be a perfect substitute for the payment d. Among other things, being a transfer d would be welfare-neutral, whereas artificially inflating the cost k would directly decrease welfare.

Alternatively, the legislator could adopt the *negligence rule*, demanding that $e \geq e^{SB}$ and using the information conveyed by the notice to check whether the platform is effectively complying with the filter requirement. *De facto* this would lead to a double standard of care: the use of an adequate filter technology that operates ex-ante and the prompt removal of content which is found to be copyright infringing ex-post, that works through the notice and take-down system.

Another approach could be that of subsidizing the investment in the filter so that the platform would find it privately optimal to set $e = e^{SB}$. The efficient implementation of this policy would suffer from the usual hurdles that affect subsidies, like the need to monitor the platform's use of the resources. By contrast, the implementation of our proposed solution would be informationally less burdensome.

Limits. A trade-off occurs when $k + d < 0$: if the payment to the state is negative and lower than the processing cost of notices incurred by the platform, a moral-hazard problem would arise in that the platform may have an incentive to induce excessive filing through effort non-modeled in our paper.

Narrow view of copyright law. A hurdle to the achievement of second best may come from a narrow interpretation of the copyright law: in designing the liability rule, the legislator should rightfully take into account the welfare of all the parties involved, that is, the platform's, the contributor's, the users', and the copyright holder's. Plausibly,

[16] The desirability of promoting online intermediaries to provide transparent information about their own and users' activities and practices has been recently highlighted by Lefouili and Madio (2022). While they argue that platforms could be exempted from liability if they fulfill some reporting obligations, we suggest that such information would be needed for the proper functioning of regulation and liability.

[17] An interesting question that goes beyond the scope of the present manuscript is how to induce online platforms to truthfully disclose this critical information.

the rule considered by the legislator would only focus on the platform's profit and the copyright holder's harm, though, and this may lead to the adoption of an inefficient filter. Formally, under this narrow view, the legislator would choose $e \in [0,1]$ to maximize the following expression:

$$aD[\mathbb{1}_{A(e)} - (1-\beta)(1-e)\eta] - (1-\beta)(1-e)\eta k - (1-\beta)(1-e)(1-\eta)H - \psi(e).$$

First-order condition yields:

$$-aD[\gamma\beta + (1-\beta)(1-\eta)] + (1-\beta)\eta k + (1-\beta)(1-\eta)H = \psi'(e^{NV}). \tag{5}$$

As illustrated in the previous subsection, the platform may not consider the harm suffered by the right-holder when there is under-detection. The platform has an incentive to invest less in the filter technology than what the legislator would like it to do. Mechanisms similar to those illustrated earlier should be adopted to align the preferences of the platform with those of the legislator. Formally, the platform would incur in an expected loss d^{NV}, whenever a notice is received, that must be set equal to $\frac{1-\eta}{\eta}H$ under this narrow view.

Discounted weight on advertising revenues. Not all the advertising revenues may be associated with informative advertising, but some may only involve a redistribution of utility from the users to the platform.[18] If so, not all advertising revenues may enter the legislator's welfare function. In particular, suppose that the legislator would only consider the fraction $\delta \in [0,1]$ of the advertising revenues that are truly welfare-increasing and this would lead to a higher expected loss d that should be imposed on the platform, for otherwise the filter would be too loose. In the following remark, we characterize d^{δ} that achieves second-best in this setting.

Remark 1. *To achieve second-best efficiency when only a fraction δ of the advertising revenues are welfare increasing, the legislator should set $d^{\delta} > d^{*}$ and equal to:*

$$d^{\delta} = \frac{(1-\eta)H}{\eta} - \frac{[V + B - (1-\delta)aD][\gamma\beta + (1-\beta)(1-\eta)]}{(1-\beta)\eta}. \tag{6}$$

Cross-group externalities. The economics literature on platforms has highlighted the role played by cross-group externalities (see, among others, Caillaud and Jullien, 2003, Rochet and Tirole, 2003, Armstrong, 2006, and Hagiu, 2006), that are absent in our baseline model. The main conclusions of our paper are not qualitatively altered if we assume that the benefits users (respectively, contributors) enjoy are increasing in the

[18] There is a heated debate over the issues inherent in the platforms' ad-funded model and some commentators have invoked the introduction of taxes on the revenue that platforms collect from digital ads (e.g., see Paul Romer's "Taxing Digital Advertising" in his own website).

number of contributors (users) the platform manages to attract. To see this, let us now make some changes to the baseline model. Assume that there is a unit mass of contributors where $\beta \in (0,1)$ now denotes the fraction of contributors who develop original content and let the users' utility be increasing in the amount of content which is made available online. As a stricter filter, i.e., a higher level of e, reduces the amount of content available on the platform, it also reduces the utility that users expect to receive from joining the platform. Consequently, we assume that both $D(e)$ and $V(e)$ are strictly decreasing in e.[19] The contributors' inalienable benefits are also assumed to be increasing in the fraction of users who join the platform, that is, $B(e)$, with $B'(e) < 0$.[20]

As we show in the appendix, in the presence of cross-group externalities the legislator should bear in mind that a stricter filter reduces the benefits that users derive from joining the platform and, in turn, this negatively affects the contributors' well-being. The second-best level of e is lower the higher the magnitude of these cross-group externalities and so is the expected loss that it should impose on the platform to align its incentives to develop a copyright filter.

Simulation. In the appendix, we provide a numerical illustration of the results obtained in this section and we compare the welfare effects of alternative policies. Specifically, we take some specific parameter values and we determine the second-best and the market-based solution as well as the size of the loss d that the legislator should impose onto the platform to restore efficiency. We find that overall welfare increases by 38% with respect to the current policy: while the right-holder is better off, the platform is hurt by the proposed policy that implements second best; moreover, users and contributors are worse off.[21]

4 Right-holder's Type-I Errors

In this section, we relax the assumption that the notice system does not entail Type-I errors, but we assume that the right-holder can invest in a technology that reduces the occurrence of such errors. Real world examples of technologies that detect potential infringements are the automated notice systems developed by large corporations to identify

[19] In the appendix, we provide an example of how a stricter filter can reduce the generic user i's utility from joining the platform, v_i, by shifting to the left the distribution of the expected benefits in a first-order stochastic dominance sense.

[20] In fact, $B(e)$ is function of the demand and not the filter itself. However, as $D(e)$ is a decreasing function of e, this simplification does not affect our results and we keep like it is for the sake of conciseness.

[21] There are two caveats that we should consider. First, the right-holder does not make any Type-I errors. Second, the users do not value the originality of the content. These assumptions are relaxed in Sections 4 and 5.1, respectively.

infringing content and send notices to platforms. When designing the policy, the legislator should also be concerned with reducing over-removal. This is crucial for the proper functioning of the first pillar of the policy, described in the previous section: the quality of the filter technology cannot be inferred by the number of notices received if there are too many type-I errors - we henceforth say that the right-holder sends an excessive number of notices when these are not all meritorious.

At the same time, the possibility of type-I errors on the right-holder's side may lead to litigation: the platform may decide not to take down automatically the content upon receiving a notice. In making this decision, the platform will take into account the costs and benefits of challenging a notice: besides the unavoidable legal expenses, the platform will have to pay damages if it loses. The upside is that the platform gets to keep the advertising revenues generated by the content if it wins the trial and may obtain some payment from the right-holder. In weighing the pros and cons of challenging a notice, the platform will consider the probability that the received notice is indeed meritorious. In turn, the right-holder may find it unprofitable to send too many notices to avoid legal expenses and possible damages that must pay if it loses when the notice is deemed groundless by the court.

The transfers that the platform and the right-holder have to exchange depending on the outcome of the trial can be thought of as two policy variables that the legislator can use to align the players' incentives. Indeed, we will show that fine-tuning their size is crucial to achieving second-best. This is because these two instruments can motivate the right-holder to invest in a better automated notice technology and discipline its incentives to send out notices.

To model the automated notice technology, we assume that the right-holder detects infringements with some exogenous probability. However, the baseline technology gives rise to type-I errors and the right-holder can invest to reduce their occurrence. Formally, the right-holder privately observes a signal $s_R \in \{0, 1\}$ and sends a notice when $s_R = 0$. The probability of detecting an infringement is $Pr[s_R = 0|\theta = 0] = \eta \in [0, 1]$. The technology may be imperfect in that $Pr[s_R = 0|\theta = 1] = (1-\gamma_R)\eta$. The probability $\eta \leq 1$ is exogenously given and the right-holder can privately invest to increase $\gamma_R \in [0, 1]$. This reduces the probability of type-I errors at cost $\phi(\gamma_R)$, which is increasing and convex in its argument.[22]

When the platform receives a notice, it bears a processing cost k and can then challenge

[22] For simplicity, we have assumed that $Pr[s_R = 0] = \eta$, irrespective of θ if the right-holder does not invest. More plausibly, in that occurrence $Pr[s_R = 0|\theta = 1] < \eta$, or else the right-holder would send many utterly meritless notices. If that were an issue, the legislator should think of penalties for sending frivolous take-down notices (in this regard, see the related penalties for frivolous lawsuits discussed by law and economics scholars, e.g., Bone, 1997).

or not the notice. If the notice is accepted, the content is removed from the platform, resulting in zero profit. As for the right-holder, if the notice is approved and the content taken down, the right-holder does not incur in any loss.

If the notice is challenged, the platform and the right-holder incur in legal expenses $l^P \geq 0$ and $l^R \geq 0$, respectively. Moreover, the right-holder suffers some harm H if the content is infringing due to the delay in the trial decision. If the platform wins the lawsuit, the content stays online and the right-holder pays $d^R \geq 0$ to the platform. Conversely, if the claim is upheld, the content is removed and the platform transfers $d^P \geq 0$ to the right-holder. We let the legislator set d^P, d^R, and the payment to the state d. We assume that there are no information gains stemming from the lawsuit, but when deliberating the court is known to be right on average.[23]

We amend the sequence of events of our game from Stage 4 onwards.

4. The right-holder makes an investment γ_R to reduce the occurrence of type-I errors in its automated notice technology. Then, it observes a signal $s_R \in \{0, 1\}$ and decides whether or not to send a notice.

5. If a notice is received, the platform decides whether to take down the content or bring it to the court. All players derive utility.

4.1 New Second-best

In this subsection, we determine the socially efficient choice of the notice technology γ_R at Stage 4, as well as the new optimal choice of e at Stage 1. The two choices will be interdependent.

First, we determine under what condition the observation of $s_R = 0$ would trigger a notice. This is the case if the harm to the right-holder associated with keeping the content online is larger than the loss of profits, users' and contributors' benefits, and the cost of processing a notice k that is incurred whenever a notice is sent, that is if

$$H\,Pr[\theta = 0|s_R = 0] = H\frac{Pr[\theta = 0|A]\eta}{Pr[A]} > aD + V + B + k.$$

Consider that the harm H is suffered only if the notice is well-founded, whereas the loss associated with content removal is independent of such probability. If this inequality does

[23] In our set-up, there are evident incentives for the right-holder and the platform to settle out of court to save the litigation costs. Indeed, Article 17 (9) of the EU Directive on Copyright in the Digital Single Market sponsors out-of-court settlements of disputes. In the analysis that follows, we abstract from this possibility but we remark here that our main conclusions would go through given the symmetric information between the two litigants and under other relatively mild assumptions (for a thorough analysis of private dispute resolution mechanisms in the shadow of litigation, see Spier, 2007).

not hold, the right-holder should not send a notice after observing $s_R = 0$ and, as a result, it would be optimal to set $\gamma_R = 0$. Instead, if the inequality holds, it would be socially desirable to choose γ_R so that original content is not removed too often. Specifically, γ_R should be chosen to maximize:

$$-[aD + V + B + k]Pr[s_R = 0] - \phi(\gamma_R),$$

where

$$Pr[s_R = 0] = Pr[\theta = 1|A](1 - \gamma_R)\eta + Pr[\theta = 0|A]\eta.$$

We obtain the best response of γ_R as function of e.

$$\phi'(\gamma_R(e)) = [aD + V + B + k]Pr[\theta = 1|A]\eta.$$

In words, a higher γ_R reduces the occurrence of type-I errors, thereby providing benefits to the platform, the contributor, and the users. To see that $\gamma_R(e)$ is increasing in e, consider that:

$$Pr[\theta = 1|A] = \frac{\beta(1 - \gamma e)}{\beta(1 - \gamma e) + (1 - \beta)(1 - e)},$$

which is increasing in e.

Supposing now that $s_R = 0$ triggers a notice, the legislator would choose $e \in [0, 1]$ to maximize the following social welfare expression:[24]

$$\begin{aligned} W^{SB} =&[aD + V + B][\mathbb{1}_{A(e)} - (1 - \beta)(1 - e)\eta - \beta(1 - \gamma e)(1 - \gamma_R(e))\eta] \\ &-[(1 - \beta)(1 - e)\eta + \beta(1 - \gamma e)(1 - \gamma_R(e))\eta]k - (1 - \beta)(1 - e)(1 - \eta)H \\ &-\mathbb{1}_{A(e)}\phi(\gamma_R(e)) - \psi(e). \end{aligned} \tag{7}$$

The key difference with respect to the welfare expression (1) in the baseline model is that we are now allowing for type-I errors in the right-holder's notice technology. This reduces the probability that content is kept online and increases the chances that the platform will have to process a take-down notice. Denote e^{SB} as the second-best level. This is derived implicitly from the following first-order condition:[25]

$$\begin{aligned} &-[aD + V + B]\left[\gamma\beta[1 - (1 - \gamma_R^{SB})\eta] + (1 - \eta)(1 - \beta)\right] \\ &+\eta\left[(1 - \beta) + \beta\gamma(1 - \gamma_R^{SB})\right]k + (1 - \eta)(1 - \beta)H + [\beta\gamma + (1 - \beta)]\phi(\gamma_R(e^{SB})) = \psi'(e^{SB}). \end{aligned} \tag{8}$$

Note that the effect of a change in e on the equilibrium value of γ_R does not appear in the above expression due to the Envelope Theorem. In itself the fact that the right-holder can make type-I errors (i.e., $\gamma_R < 1$) implies that the platform's investment in developing the filter has a relatively less important role in determining which content stays online.

[24] Importantly, note that if the right-holder does not send a notice, we retrieve the case described in the baseline model with $\eta = 0$.

[25] We assume the existence of a solution: this requires that $\psi'^{-1}(\cdot)$ takes value in $[0, 1]$. Therefore, the lhs of (8) must be positive and $\psi(\cdot)$ must be sufficiently convex.

4.2 Platform's and right-holder's incentives

In this subsection, we analyze the platform's and the right-holder's incentives to invest in the filter technology and the automated notice system, respectively, in the face of potential liability and other payments that they might have to make.

As we have seen in the previous section, when there are no type-I errors, i.e., $\gamma_R = 1$, upon receiving a notice, the platform bears the processing cost k and always removes the content to avoid incurring in legal expenses and paying damages with probability 1 (i.e., the complaint would always be sustained by the court).

Consider now the polar case in which $\gamma_R = 0$. Then, as $Pr[s_R = 0]$ is independent of θ and equal to η, it follows that $Pr[\theta = 1|s_R = 0] = Pr[\theta = 1|A]$ and $Pr[\theta = 0|s_R = 0] = Pr[\theta = 0|A]$. The platform will challenge the notice when

$$Pr[\theta = 1|A][a\,D + d^R] - Pr[\theta = 0|A]d^P > l^P, \tag{9}$$

where recall that $a\,D$ is the platform's profits if the content is kept online, d^R and d^P are the damages paid (respectively, received) by the right-holder to (from) the platform if the notice is dismissed (upheld) in court, whereas l^P are the platform's legal expenses. Notice that the platform is more likely to challenge a notice if the profit it obtains by keeping the content online is larger, the damages that the right-holder would pay and the probability that the content is indeed original are higher, the damages the platform would pay if it loses and the legal expenses are lower.

Remark 2. *If condition (9) is not satisfied, then the right-holder will never make an effort to reduce the occurrence of type-I errors.*

This result provides a cautionary tale for the legislator: increasing damages asked to the platform (and making the legal process more onerous) can reduce the likelihood that the platform challenges a notice, effectively killing any right-holder's incentive to avoid type-I errors. Its consequence would be an excessive number of notices.

More in general, for $\gamma_R \in (0, 1)$, the platform will challenge the notice in Stage 5 when the following inequality holds:

$$Pr[\theta = 1|s_R = 0][a\,D + d^R] - Pr[\theta = 0|s_R = 0]d^P > l^P, \tag{10}$$

where

$$Pr[\theta = 1|s_R = 0] = \frac{Pr[\theta = 1|A](1 - \gamma_R)\eta}{Pr[\theta = 1|A](1 - \gamma_R)\eta + Pr[\theta = 0|A]\eta};$$

$$Pr[\theta = 0|s_R = 0] = \frac{Pr[\theta = 0|A]\eta}{Pr[\theta = 1|A](1 - \gamma_R)\eta + Pr[\theta = 0|A]\eta}.$$

From (10), we can see that the platform could find it profitable to invest in a stricter filter at Stage 1, so that the content approved is more likely to be original and its willingness to challenge a notice in Stage 5 higher. That is, investing in a stricter filter can act as a commitment device for the platform, thereby prompting the right-holder to reduce the occurrence of type-I errors to avoid litigation. Moreover, it is possible to show that (10) is more likely to hold for low values of γ_R. This is expected: the platform is more willing to challenge a received notice if there is a higher probability that the right-holder makes type-I errors.

Right-holder's problem. Let us now focus attention on the right-holder's problem in Stage 4. To this end, consider the right-holder's incentives to invest in γ_R subject to the platform challenging or accepting the notice. The right-holder may decide to maximize:

$$\max_{\gamma_R \in [0,1]} \quad -(1-\eta)Pr[\theta = 0|A]H - \phi(\gamma_R),$$

subject to the platform removing the content after receiving a notice. Namely, inequality (10) must not be satisfied:

$$Pr[\theta = 1|s_R = 0][a\,D + d^R] - Pr[\theta = 0|s_R = 0]d^P \leq l^P.$$

While γ_R negatively affects the right-holder's utility, it helps satisfy the constraint. In this case, the right-holder chooses the lowest γ_R that makes the constraint bind. Specifically, we find that:[26]

$$\gamma_R^I(e) = 1 - \frac{(1-\beta)(1-e)\eta(d^P + l^P)}{\beta(1-\gamma e)\eta[aD + d^R - l^P]}. \tag{11}$$

Alternatively, the right-holder may decide to maximize:

$$\max_{\gamma_R \in [0,1]} \quad -Pr[\theta = 0|A]H - \phi(\gamma_R) + \eta Pr[\theta = 0|A](d^P - l^R) - \eta(1-\gamma_R)Pr[\theta = 1|A](d^R + l^R),$$

subject to the platform challenging the notice, i.e., inequality (10) holds. In this case, the right-holder chooses γ_R to reduce the probability that its claim is denied by the court taking into account the effort cost. This motivates the right-holder's effort to decrease type-I errors only if $d^R + l^R > 0$, that is, if there are some damages the right-holder has to pay if the notice is found to be groundless. The right-holder's best response would be:

$$\phi'(\gamma_R^{II}(e)) = \frac{\eta\beta(1-\gamma e)(d^R + l^R)}{(1-\gamma e)\beta + (1-e)(1-\beta)}. \tag{12}$$

This solution holds if, at the above $\gamma_R(e)$, inequality (10) holds. Since (10) is more likely to hold for low values of γ_R, it must be that $\gamma_R^I(e) > \gamma_R^{II}(e)$, which is more likely to be the case if legal expenses are lower. The right-holder will choose the option that gives it the highest expected payoff. Denoting by π_R the right-holder's profit, we obtain the following lemma.

[26] It requires $l^P < aD + d^R$.

Lemma 1. *The right-holder prefers to avoid litigation if* $\pi_R\Big(\gamma_R^I(e)\Big) \geq \pi_R\Big(\gamma_R^{II}(e)\Big)$, *that is, if:*

$$\eta\Big(Pr[\theta=1|A](1-\gamma_R^{II})(d^R+l^R)+Pr[\theta=0|A](H-d^P+l^R)\Big) \geq \phi(\gamma_R^I(e))-\phi(\gamma_R^{II}(e)). \quad (13)$$

While avoiding a lawsuit requires a higher effort to fine-tune the automated notice system, it is relatively more beneficial for the right-holder when the damages it would have to pay for a wrongful notice and the legal expenses are higher, and when the damages it could recover from the platform are lower. Put differently, avoiding litigation is beneficial to the right-holder if at least one of the following inequalities holds strictly: $d^P \leq H$, $d^R \geq 0$, and $l^R \geq 0$.

Platform's problem. We now go on to study the platform's investment to develop the filter technology in Stage 1. Since our aim is to determine how the legislator can achieve second-best and litigation is wasteful, below we focus on the case in which the platform removes the content upon receiving a notice. We will later pin down the policy instruments to implement this solution. The platform's maximization problem is as follows:[27]

$$\begin{aligned}\max_{e\in[0,1]} \quad & aD[\mathbb{1}_{A(e)} - (1-\beta)(1-e)\eta - \beta(1-\gamma e)(1-\gamma_R^I(e))\eta] \\ & - [(1-\beta)(1-e)\eta + \beta(1-\gamma e)(1-\gamma_R^I(e))\eta](k+d) - \psi(e),\end{aligned}$$

subject to the right-holder choosing $\gamma_R^I(e)$. This means that the right-holder is better off choosing an automated notice system that leads to the notices being accepted rather than one that results in a lawsuit. Formally, inequality (13) must be satisfied.

When the right-holder's incentive constraint is slack (i.e., inequality (13) does not bind), the first-order necessary condition yields:

$$\begin{aligned}&-aD\left[\gamma\beta[1-(1-\gamma_R^I(e^{**}))\eta] + (1-\eta)(1-\beta) - \beta(1-\gamma e^{**})\eta\frac{\partial\gamma_R^I(e^{**})}{\partial e}\right] \\ &+\left[(1-\beta)\eta + \beta\gamma(1-\gamma_R^I(e^{**})) + \beta(1-\gamma e^{**})\eta\frac{\partial\gamma_R^I(e^{**})}{\partial e}\right](k+d) = \psi'(e^{**}),\end{aligned} \quad (14)$$

where $\frac{\partial\gamma_R^I(e^{**})}{\partial e} > 0$. Because of this complementarity, the platform is more willing to invest in a stricter filter, anticipating that this will lead the right-holder to make a higher effort to avoid sending excessive mistaken notices. This increases the probability that the platform can obtain the profits associated with posting content and avert the notice processing cost k. The following remark summarizes the equilibrium pair (e^{**}, γ_R).

[27] Consistently with what we did in the baseline model, we define d as the expected loss suffered by the platform when a notice is received, in addition to k.

Remark 3. *Suppose the right-holder can make type-I errors. If at e^{**}, constraint (13) is satisfied, then the solution is given by the system made up of e^{**}, determined from (14), and $\gamma_R = \gamma_R^I$, determined from (11). If at e^{**}, constraint (13) is not satisfied, then the solution is given by the system made up of e such that (13) binds, and $\gamma_R = \gamma_R^I$, determined from (11).*

4.3 Implementing Second-Best

In this subsection, we seek the values of the damages that can implement the second-best solution. To obtain second best, the legislator needs to impose that $\gamma_R^I = \gamma_R^{SB}$ and, simultaneously, $e^{**} = e^{SB}$. The legislator's instruments are d^P, d^R, and d. In choosing them, the legislator must also ensure that the platform does not find it profitable to induce a lawsuit by choosing a too lax filter that results in γ_R^{II}.

It is convenient to use d^P and d^R to discipline the right-holder's incentive to make an effort to reduce type-I errors. As stated earlier, this requires that the platform challenge a notice if the probability of type-I errors is sufficiently high. So, it must be that the threat of litigation is severe enough that the right-holder sets γ_R as high as γ_R^{SB}. When these instruments induce the desired right-holder's behavior, the role of the notices as an informative signal of the quality of the platform's filter technology is restored and, accordingly, the legislator could set a payment d to induce the platform to choose $e = e^{SB}$, in a fashion similar to that described in the previous section. In the following proposition, we illustrate the policy (d^P, d^R, d^{**}) that induces second-best.[28]

Proposition 2. *When the right-holder can make type-I errors, to achieve second-best the legislator must set:*

1. *d^P and d^R in such a way that the right-holder would choose $\gamma_R = \gamma_R^I = \gamma^{SB}$ when $e = e^{SB}$ and avoid litigation, i.e., (13) must hold for all e;*

2. *d in such a way that (14) and (8) coincide:*

$$\begin{aligned} d^{**} =& \frac{(1-\beta)(1-\eta)H - [V+B]\left[\gamma\beta[1-(1-\gamma_R^{SB})\eta] + (1-\eta)(1-\beta)\right]}{(1-\beta)\eta + \gamma\beta(1-\gamma_R(e^{SB}))\eta + \beta(1-\gamma e^{SB})\eta\frac{\partial\gamma_R(e^{SB})}{\partial e}} \\ &- \frac{(aD+k)\left[\beta(1-\gamma e^{SB})\eta\frac{\partial\gamma_R^{SB}}{\partial e}\right] + [\gamma\beta + (1-\beta)]\phi(\gamma_R^{SB})}{(1-\beta)\eta + \gamma\beta(1-\gamma_R(e^{SB}))\eta + \beta(1-\gamma e^{SB})\eta\frac{\partial\gamma_R(e^{SB})}{\partial e}}. \end{aligned} \tag{15}$$

As the proposition states in point (1), the legislator sets d^P and d^R to solve (13)

[28] We denote the payment the platform must make by d^{**} to distinguish it from d^* of the previous section.

together with the following implicit equation:

$$1 - \frac{(1-\beta)(1-e^{SB})\eta(d^P + l^P)}{\beta(1-\gamma e^{SB})\eta[aD + d^R - l^P]} = \phi'^{-1}\left(\frac{\beta(1-\gamma e^{SB})\eta[aD + V + B + k]}{\beta(1-\gamma e^{SB}) + (1-e^{SB})(1-\beta)}\right).$$

From the above condition, it is possible to see that if the legislator wants to increase d^P, it must also increase d^R. Importantly, setting $d^P = H$ and $d^R = 0$ is unlikely to be a solution. Constraint (13) ensures that the right-holder wants to avoid litigation. From this condition, it is possible to see that d^P and d^R have counteracting effects. Firstly, an increase in d^P strengthens the right-holder's expected gain from litigation, while it reduces the platform's incentive to challenge the notice. This implies that a lower level of γ_R is needed to ensure that the notice is accepted by the platform. An increase in d^R makes the platform more willing to challenge a notice for any given frequency of type-I errors. Therefore, ensuring that there is no litigation is more demanding for the right-holder. At the same time, an increase in d^R may not help satisfy (13) because it also reduces the right-holder's utility of going to trial: were the notice dismissed in court, the right-holder would have to make a larger transfer to the platform. The bottom line is that the legislator cannot set d^P independently of d^R and imposing higher damages on the platform when the court finds that there has been a copyright infringement calls for setting higher damages on the right-holder when the claim is not upheld in court.

To summarize, a way to implement second-best, which includes (i) an adequate filter technology by the platform, to avoid that there is too much copy-right infringing content online, and (ii) an automated notice technology by the right-holder that does not lead to content over-removal, consists of the following three pillars:

1. The payment that the platform could obtain from the right-holder (or, to put it differently, the damages the platform could recover) if there is litigation cannot be too small. This is the variable d^R in our model. By doing so, the platform would be willing to challenge dubious notices in court, and this is necessary to provide the right-holder with incentives to reduce type-I errors.

2. On the other hand, the damages the right-holder could recover by going to court (denoted d^P in the model) cannot be too large relative to d^R. In other words, the two instruments are not independent of one another. If d^R is small, it is better to set d^P below the true harm H. If $d^P = H$, d^R cannot be too small. This pair strengthens the right-holder's incentives to make an effort to avoid type-I errors since the right-holder does not want too many controversies to arise.

3. The first two pillars ensure that the number of notices is not excessive and, consequently, there is no over-removal ex-post. This implies that the number of received and accepted notices can be used as a signal of the quality of the platform's filter.

> Then, to provide the platform with incentives to develop an adequate filter, the legislator could think of a payment to the state or fine the platform has to make if relatively too many notices are received, similarly to what described in the previous section. When the right-holder can make type-I errors, this payment must be amended and is given by d^{**} in the model.

In practice, there may be institutional hurdles to imposing large transfers on the right-holders when they claim their copyrights, but they lose in court, i.e., d^R may be at most very small. Under this institutional limitation, our model shows that an OHP should not face onerous damages if it refuses to take down content following a notice that holds up in court. This is especially the case when an argument for fair use of copyright-protected material can in principle be made.

In Figure 2, we graphically represent the interplay between d^P and d^R.[29] If the legislator wants to set no payment from the right-holder to the platform if the notice is dismissed in court, then d^P must be somewhat smaller than the actual disutility that the right-holder suffered because of the copyright infringement. In the case depicted in the figure, when $d^R = 0$, $d^P = 11.8$, that is, a little more than one hundredth of the disutility $H = 1,000$. If the legislator wants to make the right-holder whole should the notice be upheld in court (that is, if $d^P = H = 1,000$), then also the payment that the right-holder should make if the notice is dismissed should be sizable (i.e., $d^R = 670.5$).

Narrow view. Akin to the previous section, the legislator may pursue a narrow interpretation of the copyright law, where only the platform's profit and the right-holder's utility are taken into account. In that case, the narrow-view policy differs from the optimal one, characterized in Proposition 2 in that the terms that depend on V and B disappear. This would translate into a higher expected loss d^{**} to the platform.

5 Extensions and Robustness Checks

In this section, we extend the baseline model of Section 3 in several directions to verify the robustness of our policy prescriptions.

5.1 Intrinsic value of original content

To simplify the analysis, in the baseline model we have abstracted away from any value that users may attach to the originality of the content. Therefore, the only reason why it

[29] We take the same parameter values as for Figure 1 (see Appendix 6), with the exception of $\psi(e)$ which is now equal to $12.5e^2$. In addition, we set $\gamma = 0.2$ and $\phi(\gamma_R) = 12.5\gamma_R^2$.

may be desirable that the platform invest in the filter technology is to avoid the negative externality suffered by the right-holder when the posted content is unoriginal. Relaxing this assumption would not alter the gist of our policy prescriptions. However, the legislator would have to take into account the added value that original content brings about when setting the policy.

As users are now assumed to attach some value to the originality of the content, we amend their utility in the following way:

$$V_i = [v_i + v\theta - p]\mathbb{1}_A,$$

where $v \in [0, |\underline{v}|)$ represents the benefit from enjoying original content. In stage 3, a user will join the platform so long as she expects to receive non-negative utility given the belief on the contributor's type. We denote this belief by θ^E. The user who is indifferent between joining or not the platform is the one whose expected utility satisfies:

$$v_i = -v\theta^E + p.$$

In what follows, we continue to assume that a is sufficiently high so that the platform does not set a positive price. A notable difference with the baseline model is that, even in the absence of liability, the platform might have an incentive to invest in the filter to boost its demand. To see this, consider the platform's maximization problem in the absence of liability:

$$\max_{e} \quad aD(e)A(e) - \psi(e),$$

where $A(e) = \mathbb{1}_A$ and $D(e) = 1 - F(-v\theta^E)$, which is increasing in θ^E. To evaluate whether or not to join the platform, the users will consider the posterior probability that the content is original given that it has been approved. Using the Bayes' rule:

$$\theta^E = Pr[\theta = 1|A] = \frac{Pr[\theta = 1]Pr[A|\theta = 1]}{Pr[A]} = \frac{\beta(1 - \gamma e)}{(1 - \gamma e)\beta + (1 - e)(1 - \beta)}.$$

The first-order condition from which we derive the platform's choice of e yields:

$$aD'(e)A(e) + aD(e)A'(e) = \psi'(e).$$

The first term on the left-hand side, $aD'(e)A(e)$ represents the demand-boosting effect of marginally tightening the filter. The second term $aD(e)A'(e)$ is the downside of increasing e: this makes it less likely that some content is available online. The term on the right-hand side is the direct cost of setting a stricter filter. For an interior solution, the first-order condition can be rewritten as:[30]

$$a\left[\frac{f(-v\theta^E)v\beta(1 - \beta)(1 - \gamma)}{A(e)} - (1 - F(-v\theta^E))[\gamma\beta + (1 - \beta)]\right] = \psi'(e). \tag{16}$$

[30] The left-hand side is weakly decreasing in e if F is weakly convex, as in the case of the Uniform distribution.

There is an interior solution only if the demand boosting effect of tightening the filter outweighs the negative effect due to the reduction in the amount of content posted online.

Remark 4. *In the absence of liability, the platform will invest in the filter only if the following inequality is satisfied:*

$$v > \left[\frac{1 - F(-v\theta^E)}{f(-v\theta^E)}\right]\left[\frac{[\gamma\beta + (1-\beta)]A}{\beta(1-\beta)(1-\gamma)}\right].$$[31]

Crucially, e derived from the first-order condition is an increasing function of v, i.e., the weight attached to original content by users. That is, $e'(v) \geq 0$ and $e(v = 0) = 0$.

Adding $p > 0$ does not affect qualitatively the results. In fact, a positive fee would strengthen the case for tightening the filter: not only can the platform increase demand and boost its advertising revenue, but it can also impose a higher fee to users.

In second best, the social planner would choose $e \in [0, 1]$ to maximize the following social welfare expression:

$$\begin{aligned} W^{SB} =& [aD(e) + V(e) + B][\mathbb{1}_{A(e)} - (1-e)\eta(1-\beta)] \\ & -(1-e)\eta(1-\beta)k - (1-e)(1-\eta)(1-\beta)H - \beta c - \psi(e). \end{aligned}$$

Consider that:

$$D(e) = \int_{-v\theta^E}^{\overline{v}} f(x)dx,$$

whereas

$$V(e) = \int_{-v\theta^E}^{\overline{v}} F(x)dx + v\theta^E \int_{-v\theta^E}^{\overline{v}} f(x)dx.$$

Denote e^{SB} as the second best level. To achieve second-best, the platform must be induced to take into account the surplus that users obtain when they enjoy original content. The next proposition characterizes the expected loss that the legislator should impose on the platform when users attach value to the originality of the content.

Proposition 3. *To achieve second-best efficiency, the legislator should set d^{IV} in such a way that the platform internalizes that users attach value to the originality of the content:*

$$\begin{aligned} d^{IV} =& \frac{(1-\eta)H}{\eta} - \frac{[V(e^{SB}) + B][\gamma\beta + (1-\eta)(1-\beta)]}{(1-\beta)\eta} \\ & + \frac{V'(e^{SB})[A(e^{SB}) - (1-\beta)(1-e^{SB})\eta]}{(1-\beta)\eta}. \end{aligned}$$

[31] In the case of the uniform distribution, it is possible to show that this inequality becomes:

$$v > \frac{\overline{v}[\gamma\beta + (1-\beta)]A}{\beta[(1-\beta)(1-\gamma) + (1-\gamma e)[\gamma\beta + (1-\beta)]]}.$$

Besides the term in the first line V being now a function of e, the chief difference with expression (4) in Proposition 1 is the term in the second line. This term is positive and meant to induce the platform to internalize the users' benefits from enjoying original content. The more the users value such original content, the larger the expected loss d the platform should incur when a take-down notice is accepted, since these signals that unoriginal content was approved.

5.2 Delegation of the copyright filter

Although some platforms have developed their copyright filters in-house, like YouTube with ContentID, other platforms have decided to adopt automatic content recognition (ACR) softwares provided by third parties. The most notable example is that of Audible Magic's ACR service which is used by social networks and online sharing platforms such as Facebook, Twitch, and Vimeo. In this subsection, we investigate how the delegation of the copyright filter affects our mechanism. The platform will maintain liability for copyright infringement as it has no legal ground to make a claim against the filter developer.

We amend the baseline model by assuming that the platform pays a transfer $T \geq 0$ to a wealth-constrained filter developer that furnishes the filter in stage 1. For simplicity, all the bargaining power resides with the platform and we use the same notation as in the previous sections to denote the filter developer's investment cost function. We begin by considering the simplest scenario wherein e is contractible. The platform chooses e and T to maximize the following objective function:

$$aD[\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]-(1-\beta)(1-e)\eta(k+d)-T,$$

subject to the filter developer's participation constraint: $T-\psi(e)\geq 0$. As the platform leaves no rent to the filter developer, $T=\psi(e)$ and the maximization problem coincides with the one illustrated in Section 3.

If the strictness of the filter is non-verifiable, the platform must design an incentive contract to induce the filter developer to choose the desired level of e. Intuitively, the transfer T will be paid whenever some content is flagged by the filter as infringing. The platform's maximization problem is as follows:

$$\max_{e\in[0,1],T\geq 0} \quad aD[\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]-(1-\beta)(1-e)\eta(k+d)-[(1-\beta)e+\beta\gamma e]T,$$

subject to the filter developer's incentive compatibility and participation constraints, respectively:

$$e\in\arg\max_{e'\in[0,1]} \quad [(1-\beta)e'+\beta\gamma e']T-\psi(e'), \tag{FDIC}$$

$$[(1-\beta)e+\beta\gamma e]T-\psi(e)\geq 0. \tag{FDPC}$$

The incentive compatibility constraint guarantees that the filter developer finds in its best interest to choose $e' = e$.

In the next proposition, we show that a welfare-maximizing legislator should impose a larger expected loss d when the platform delegates the development of the copyright filter to a third party than when it develops the filter in-house.

Proposition 4. *To achieve second best efficiency when the platform delegates the filter technology to a filter developer, the legislator should set $d^{Del} > d^*$.*

Because of the moral-hazard problem, the platform must give up an information rent to the filter developer. This information rent is paid up more often when e is higher. Consequently, the platform has an incentive to implement a less strict filter. To counterbalance this incentive and restore efficiency, the legislator must adjust upward the expected loss d imposed onto the platform.

We conclude by making two observations. First, for a given transfer T, the filter developer may have an incentive to inflate the frequency of type-I errors, i.e., by increasing γ. By doing so, the filter developer would receive the payment T more often, but too little content would be made available on the platform. To overcome this issue, the platform may offer a profit-sharing agreement whereby the filter developer receives a transfer also when some content is approved and is not successively taken down by the platform following a notice. Second, if all the bargaining power resides with the filter developer and this can extract the entire platform's surplus, then there is no difference between d^{Del} and d^*. In practice, the most established filter developer normally charges platforms an upfront price and a fee for each file analyzed (i.e., a two-part tariff) which may enable it to extract their surplus.

5.3 Platform's type-I errors

In this subsection, we assume that the platform can also affect the frequency of type-I errors, in addition to the strictness of the filter, e. We suppose that there is an initial level of γ, denoted $\overline{\gamma} \in (0, 1]$, and the platform can invest to reduce it by choosing $\gamma_1 \in [0, \overline{\gamma}]$, so that the final $\gamma := \overline{\gamma} - \gamma_1$. Specifying the filter technology costs: $\psi(e, \gamma_1)$, with $\psi_e(e, \gamma_1) > 0$, $\psi_{\gamma_1}(e, \gamma_1) > 0$, $\psi_{ee}(e, \gamma_1) > 0$, $\psi_{\gamma_1\gamma_1}(e, \gamma_1) > 0$, $\psi_{ee}(e, \gamma_1)\psi_{\gamma_1\gamma_1}(e, \gamma_1) - 2\psi_{e\gamma_1}(e, \gamma_1) > 0$. Moreover, $\psi(1, \cdot) = \infty$ and $\psi(e, \overline{\gamma}_1) = \infty$ for any $e > 0$. In second-best, the social planner would choose the level of e and γ_1 that satisfy the following two first-order conditions:

$$\begin{aligned} -[aD + V + B][(\overline{\gamma} - \gamma_1^{SB})\beta + (1 - \beta)(1 - \eta)] + (1 - \beta)[\eta k + (1 - \eta)H] =& \psi_e(e^{SB}, \gamma_1^{SB}); \\ [aD + V + B]\beta e^{SB} =& \psi_{\gamma_1}(e^{SB}, \gamma_1^{SB}). \end{aligned}$$

There is an obvious complementarity between the second-best levels γ_1 and e: (i) a stricter filter makes it more critical to devote resources to reduce the occurrence of type

I-errors; (ii) likewise, reducing γ decreases the downside of adopting a stricter filter. In maximizing welfare, the legislator has to take into account that the expected loss d will affect the platform's incentives to choose both e and γ_1. The next remark illustrates the main result of this analysis.

Remark 5. *To maximize welfare, the legislator should set d such that $e(d) > e^{SB}$.*

As the legislator can only use one instrument to govern the platform's incentives to choose e and γ, welfare will fall short of second-best. Note that the legislator could always set d that induces e^{SB}. However, this is inefficient as it would lead to too many type-I errors. Since the platform's choice of γ_1 is increasing in the strictness of the filter e, to reduce the occurrence of type-I errors, the legislator will set d to induce the platform to choose a filter that is stricter than the one of second-best. Differently, if the legislator adopts a narrow view of the copyright law, setting $d^{NV} = \frac{1-\eta}{\eta}H$ continues to hold true. This is because the users' and contributors' benefits are disregarded by the legislator and, therefore, the desired level of γ_1 is achieved when e^{NV} is implemented.

5.4 Minor infringements and content monetization

In this subsection, we relax the assumption that the copyright infringement would always cause a substantial harm to the right-holder. In some occurrences, the infringement would have a limited impact on the right-holder, which is dwarfed by the benefits the content generates to the contributor, the platform, and the users. Being aware of this possibility, platforms give right-holders the option to monetize content that is found to be infringing.[32] For instance, this is the case with YouTube's Content ID where material flagged as infringing can be either blocked or monetized by the harmed right-holder, as shown in Figure 3 in the Appendix.[33,34]

In this section, we assume that H is distributed according to the continuous function $G(\cdot)$ on $[0, \overline{H}]$, where $\overline{H} > aD + V + B + k$. We assume that the platform and the right-holder do not observe freely the originality of the content, but if some content is deemed unoriginal, they both learn its expected harm. In this extension, we assume that B can be transferred from the contributor to the right-holder. Thus, the right-holder would first expropriate the contributor's benefit B. If this is not enough, it will then extract the

[32] Monetization can be seen as an effective use of consumer control of their data and this is proved to be welfare enhancing relative to both perfect price discrimination and uniform pricing (see Ali et al., 2022). On the subject of monetization, see also García (2020).

[33] Likewise, when a right-holder detects infringing content on YouTube can request either its removal or the transfer of its ownership.

[34] The same figure can be found in Madiega (2020). The source is Google.

platform's profit aD. Should this be insufficient to satisfy its claim, the right-holder will request the removal of the content from the platform.

We focus our analysis on the scenario in which the right-holder does not make I-type errors when sending notices, as in Section 3. We begin by adopting the social planner's perspective and then we explore the optimal design of liability. We notice that the social planner would post content online if it is either original, which has probability β, or unoriginal but not very harmful, which has probability $(1-\beta)G(aD+V+B)$. Accordingly, the social planner's copyright filter would allow content with the following probability:

$$\begin{aligned}\mathbb{1}_{A(e,H)} := \quad & \beta(1-\gamma e)+(1-\beta)(1-e) \\ & +[\beta\gamma e+(1-\beta)e]G(aD+V+B).\end{aligned}$$

In second best, the social planner would choose $e \in [0,1]$ to maximize the following social welfare expression:

$$\begin{aligned}W^{SB} =& [aD+V+B][\mathbb{1}_{A(e,H)}-(1-\beta)(1-G(aD+V+B+k))(1-e)\eta] \\ & -(1-\beta)(1-G(aD+V+B+k))(1-e)\eta k-(1-\beta)(1-e)(1-\eta)\int_{aD+B+V+k}^{\overline{H}} HdG(H) \\ & -(1-\beta)\int_{0}^{aD+V+B+k} HdG(H)-\psi(e).\end{aligned}$$

The first line is the overall benefit associated with approving content $(aD+B+V)$ multiplied by the probability that the content is accepted and remains online. The second line is the cost due to the copyright filter allowing very harmful content online, that is, unoriginal content with $H > aD+V+B+k$: it will either be detected by the right-holder, and costly removed by the platform, or will go undetected.[35] The third line is the cost associated with allowing unoriginal content that entails a minor harm for the right-holder, as well as the cost of developing the filter technology. At an interior solution, the optimal investment in the filter satisfies:

$$\begin{aligned}&-[aD+V+B]\Big\{(1-G(aD+V+B))[\gamma\beta+(1-\beta)]-(1-G(aD+V+B+k)(1-\beta)\eta\Big\} \\ &+(1-G(aD+V+B+k))(1-\beta)\eta k+(1-\beta)(1-\eta)\int_{aD+B+V+k}^{\overline{H}} HdG(H)=\psi'(e^{SB}(H)).\end{aligned}$$

Parties' incentives. The right-holder would always send a notice whenever content is unoriginal, requesting either the content removal or its monetization. Infringing content is kept online only if $H \leq aD+B$ as the right-holder cannot extract the users' surplus. In writing the platform's investment problem, it is helpful to distinguish between the case

[35] Recall that at this stage the content should optimally be removed if H is greater than the sum of the benefits associated with keeping the content online as well as the cost of processing the notice, k.

in which the content is original and the one in which it is not.

$$\begin{aligned}\max_{e\in[0,1]} \quad & \beta\left[(1-\gamma e)aD+\gamma e\left(G(B)aD+\int_B^{aD+B}(aD+B-H)dG(H)\right)\right]\\ & +(1-\beta)\left[(1-e)(1-\eta)aD+(e+(1-e)\eta)\left(G(B)aD+\int_B^{aD+B}(aD+B-H)dG(H)\right)\right]\\ & -(1-\beta)(1-e)\eta k-\psi(e).\end{aligned}$$

To understand the above expression, first note that the platform always obtains aD when the copyright filter approves the content and the right-holder does not detect an infringement. Moreover, the platform fully enjoys aD when either its filter or the right-holder detect an infringement whose harm is lower than the contributor's transferable benefit B. By contrast, if either the platform or the right-holder detects some infringing content and the harm is greater than B but lower than $aD+B$, the content will be kept online, but the platform's profit will be shared with the right-holder. The platform anticipates that the right-holder will send a notice whenever it detects an infringement. When this occurs, the platform bears a processing cost equal to k. At an interior optimum, the first-order condition gives:

$$-[\gamma\beta+(1-\beta)(1-\eta)]\left[(1-G(aD+B))aD+\int_B^{aD+B}(H-B)dG(H)\right]+(1-\beta)\eta k=\psi'(e^{**}(H)).$$

A noteworthy difference with the baseline model is that the copyright filter alone cannot achieve second best and the reason is twofold. First, the right-holder would send notices whenever it detects some unoriginal content, irrespective of the harm this causes. This leads to excessive processing costs. Second, because of the inability to extract the users' benefits, some unoriginal content that should optimally be kept online will be removed. This seems to suggest that the platform's ability to extract the users' benefit could increase the scope for monetization, resulting in more content being available online. Taking into account these limits, the social planner could hope for third best. We define welfare in third best as follows:

$$\begin{aligned}W^{TB} =&[aD+V+B][\mathbb{1}_{\tilde{A}(e,H)}-(1-\beta)(1-G(aD+B))(1-e)\eta]\\ & -(1-\beta)(1-e)\eta k-(1-\beta)(1-e)(1-\eta)\int_{aD+B}^{\overline{H}}HdG(H)-(1-\beta)\int_0^{aD+B}HdG(H)-\psi(e),\end{aligned}$$

where

$$\mathbb{1}_{\tilde{A}(e,H)}:=\beta(1-\gamma e)+(1-\beta)(1-e)+[\beta\gamma e+(1-\beta)e]G(aD+B).$$

At an interior optimum, the level of investment in the copyright filter that maximizes third-best welfare is:

$$\begin{aligned}&-[aD+V+B](1-G(aD+B))[\gamma\beta+(1-\beta)(1-\eta)]\\ &+(1-\beta)\eta k+(1-\beta)(1-\eta)\int_{aD+B}^{\overline{H}}HdG(H)=\psi'(e^{TB}(H)).\end{aligned}$$

Following the same approach as in Proposition 1, the social planner should impose the following expected loss onto the platform whenever a notice is received and accepted to achieve third-best:

$$d^{TB} = \frac{(1-\eta)\int_{aD+B}^{\overline{H}} H dG(H)}{\eta} - \frac{(V+B)[\gamma\beta + (1-\beta)(1-\eta)][1-G(aD+B)]}{(1-\beta)\eta} + \frac{[\gamma\beta + (1-\beta)(1-\eta)]\int_{B}^{aD+B}(H-B)dG(H)}{(1-\beta)\eta}.$$

5.5 Low advertising revenues and market power

Throughout the analysis, we have maintained the assumption that advertising revenues are so high that the platform finds it profitable to set the lowest admissible fee, i.e., $p = 0$. In fact, existing platforms tend to charge positive subscription fees to users or they let users self-select by providing a menu of subscription packages.[36] In our model, if advertising revenues are not large enough, the platform would set a positive fee for accessing its service, thereby curtailing users' demand. This also implies that the platform extracts some part of the users' surplus V, which more than compensates for the reduction in the demand. As long as users do not care about the originality of the content, the platform has an incentive to invest less in the filter technology the larger the size of their surplus that it appropriates. Intuitively, by tightening the filter, the probability that the content is available and that surplus is generated and enjoyed by the platform goes down. This translates into a higher expected loss that the legislator should impose onto the platform to align its investment incentives. Mathematically, only a fraction of the users' surplus V would enter the expression of the expected loss in (4). When the platform enjoys more market power (e.g., because there are fewer competing platforms or the users' demand is less sensitive to changes in the fee), the platform appropriates a higher fraction of the users' surplus and invests less in the filter technology. Therefore, a higher market power appears to increase the expected loss that the legislator must impose onto the platform to provide second-best incentives for the development of the filter technology.

[36] See Carroni and Paolini (2020), for a model that rationalizes the advertising-based and subscription-based business models adopted by streaming platforms. For an interesting analysis of media platforms' business models, we would also like to draw the readers' attention to the paper by Casner and Teh (2023) in which the authors consider three different business models: (i) *pure discovery mode* that enables consumers to discover content posted by creators; (ii) *pure membership mode* that allows creators to monetize on their relationships with viewers; (iii) *hybrid mode* that combines both.

6 Concluding Remarks

This paper has developed a theoretical framework that allows considering the trade-off that an online platform faces when it hosts material that could be copyright infringing. The paper has also proposed a policy to induce the platform not to exclude excessive content and, at the same time, to discipline right-holders so that they properly consider fair use when submitting take-down notices. The policy includes a pecuniary fine to the platform whenever a notice holds up or if the ratio of the accepted notices over the overall hosted content exceeds a pre-determined threshold. The higher the threshold the larger the fine that should be imposed to reach efficiency. Additionally, the damage that the right-holders could recover from going to court cannot be too large as compared to the payment they should make to the platform if they lose. As we have shown, this policy may be fine-tuned depending on, among other things, the utility users attach to enjoy original content, the size of the cross-group externalities, and whether the filter is developed in house or provided by a third party. More in general, our policy prescriptions may vary depending on the characteristics of the right-holder and the platform. In the model, we have assumed only a single type of each. However, it stands to reason that better-established right-holders are more capable of sending notices and their cost of reducing type-I errors is smaller. The platform's market power should be also taken into account in tailoring regulation and liability, as this affects its incentives and the availability of financial resources to develop a stricter filter.

In the remainder, we offer some remarks on some assumptions that could be relaxed. First, we have assumed that developing original content only depends on the contributor's ability and it does not entail large costs. In addition to adverse selection, there might also be moral hazard, e.g., the high-ability contributor may have to invest to develop original content or may decide to mimic the low-ability one by taking off-the-shelf content. This would be the case if the contributor is not intrinsically motivated to develop original content or if he has no other incentives such as building a personal reputation on the platform (Fromer, 2012). Absent those incentives, the platform might have to motivate original content creation by giving up a positive fraction of the content-generated revenues to the contributor.[37] This additional friction would not qualitatively alter our results.

Second, in the model we have assumed that a right-holder suffers some fixed harm denoted H if the posted content infringes its copyright. In reality, the size of the harm may plausibly be related to the platform's demand. This is because the platform's users would not consume the copyright-protected content outside the platform. Namely, there could be substitution between the contributor's unoriginal content available on the platform and

[37] Revenue-sharing agreements with successful content developers are common, for instance, at YouTube through its Partner Program.

the right-holder's own material. Making the right-holder's harm an increasing function of the platform's demand would not qualitatively change our results, although the optimal copyright filter would be stricter when there is a larger demand, since the expected harm suffered by the copyright holder would be greater.

In addition to the harm directly suffered by the right-holder, whose content has been unlawfully posted on the platform by a contributor, there might also be a negative externality that unoriginal content engenders and the legislator cares about. To be more specific, it is widely believed that, if intellectual property is not properly protected, artists may be discouraged from devoting resources to original content creation. Intellectual property rights bring about exclusivity and control by which creators may prevent free-riding, charge supracompetitive prices, and recoup their initial investments in developing new content (Landes et al., 2003, page 40). In order to reinforce intellectual property rights and furnish strong incentives to authors to develop their works in the first place, the optimal copyright filter would be stricter than the one reported in the text and this would call for a larger expected loss that should be imposed on the platform when it hosts copyright-infringing content. Nonetheless, some caveats should be introduced. First, creators may have other incentives that may adequately spur creativity (Fromer, 2012). Second, some creative content - think for instance of memes on social media platforms - would involve very low initial investment and therefore strong intellectual property would not be necessary. Third, intellectual property rights also entail social costs. In particular, strong intellectual property protection generates a dynamic inefficiency since downstream creators would have fewer possibilities of reusing or building on previous content (Menell and Scotchmer, 2019). The legislator should also care about these social costs when regulating copyright filters. Discussion of these issues goes beyond the scope of this article.

Acknowledgements

We thank Luis Aguiar, Yonathan Arbel, Elpiniki Bakaouka, Antonio Cabrales, Sheila Canudas, Elias Carroni, Rosa Ferrer, Brian L. Frye, Juan José Ganuza, Nuno Garoupa, Gerard Llobet, Leonardo Madio, Bertin Martens, Giuseppe Pignataro, Martin Peitz, Carlo Reggiani, Patrick Rey, Frank Schlutter, Maciej Sobolewski, Marco A. Schwarz, and Kathryn E. Spier as well as the audiences at the 38th Annual Conference of the European Association of Law and Economics (EALE), at the joint Digital Economics Seminar (jDES), at the 4th UNIBG Industrial Organization Winter Symposium, at the ZiF Research Group Workshop "Economic and Legal Challenges in the Advent of Smart Products" (University of Bielefeld), at the 2022 MaCCi Annual Conference, at the 2022 Oligo Workshop, at the JRC-TSE Workshop on "Liability in the Digital Economy"

(Seville), at the 2022 Jornadas de Economia Industrial (Universidad de Las Palmas), and at the Universitat de les Illes Balears for very useful comments and discussions. We acknowledge the generous support of the NET Institute through their 2021 Summer Grant. Alessandro De Chiara and Ester Manna also acknowledge the financial support of Ministerio de Ciencia, Innovación y Universidades through grants PID2020-114040RB-I00 and PID2021-128237OB-I00, the Ministerio de Economía y Competitividad and Fondo Europeo de Desarrollo Regional through grant ECO2016-78991-R (MINECO/FEDER, UE), and the Government of Catalonia through grant 2014SGR493. Antoni Rubí-Puig also acknowledges the financial support of Ministerio de Ciencia and Innovación through grant PID2021-126354OB-I00. The usual disclaimer applies.

Appendix A

Proof of Proposition 1

Note that the first-order condition of the platform's new maximization problem yields:

$$-aD[\gamma\beta + (1-\beta)(1-\eta)] + (1-\beta)\eta(k+d) = \psi'(e^*). \tag{A1}$$

To achieve second best, it suffices to set d in such a way that (A1) coincides with (2), which yields (4). □

Proof of Remark 1

The platform's first-order condition coincides with (A1). However, when the legislator attaches a weight δ to the advertising revenues aD, the second-best will be as follows:

$$\begin{aligned}&-[\delta aD + V + B][\gamma\beta + (1-\beta)(1-\eta)]\\&+(1-\beta)\eta k + (1-\beta)(1-\eta)H = \psi'(e^{SB\delta}).\end{aligned}$$

Then, to achieve second best, d^δ must be set in such a way that (A1) coincides with the above expression. □

Cross-group externalities

Example. Below we provide an example of how e can directly affect the users' benefits from joining the platform by reducing the available content. Let the generic user's utility v_i be distributed on $[\underline{v}, \bar{v}]$ according to the distribution $F[v|e]$ which is twice continuously differentiable with respect to e and such that $F_e[\cdot|e] > 0$ and $F_{ee}[\cdot|e] > 0$ for all v. A justification for this assumption is that an increase in e reduces the probability that content is made available, $Pr[A]$, thereby reducing the benefit that users expect to receive. Therefore, a less strict filter increases the benefits from joining the platform, by shifting the distribution of the expected benefits in a first-order stochastic dominance sense. Under the assumption that $p = 0$, the platform's demand is:

$$D(e) := \int_0^{\bar{v}} F_x(x|e)dx;$$

and the users' overall utility is:

$$V(e) := \int_0^{\bar{v}} F(x|e)dx.$$

Analysis. The legislator would choose e to maximize the following welfare function:

$$\begin{aligned}&[aD(e)+V(e)+B(e)][\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]\\&-(1-\beta)(1-e)\eta k-(1-\beta)(1-e)(1-\eta)H-\psi(e).\end{aligned}$$

The first-order condition yields the second-best level of e, denoted e^{cge}:

$$\begin{aligned}&-[aD(e^{cge})+V(e^{cge})+B(e^{cge})][\gamma\beta+(1-\beta)(1-\eta)]\\&+[aD'(e^{cge})+V'(e^{cge})+B'(e^{cge})][\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]\\&+(1-\beta)\eta k+(1-\beta)(1-\eta)H-\psi'(e^{cge})=0.\end{aligned}\tag{A2}$$

In choosing the strictness of the filter the platform would maximize the following objective function:

$$aD(e)[\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]-(1-\beta)(1-e)\eta(k+d)-\psi(e),$$

The following proposition illustrates how the legislator should set the expected loss in the presence of cross-group externalities.

Proposition 5. *To achieve second-best efficiency, the legislator should set:*

$$\begin{aligned}d^{cge}=&\frac{(1-\eta)H}{\eta}-\frac{(V(e^{cge})+B(e^{cge}))[\gamma\beta+(1-\beta)(1-\eta)]}{(1-\beta)\eta}\\&+\frac{(V'(e^{cge})+B'(e^{cge}))[\beta(1-\gamma e^{cge})+(1-\beta)(1-e^{cge})(1-\eta)]}{(1-\beta)\eta}.\end{aligned}\tag{A3}$$

Proof. Note that the first-order condition of the platform's maximization problem yields:

$$aD'(e)[\beta(1-\gamma e)+(1-\beta)(1-e)(1-\eta)]-aD(e)[\beta\gamma+(1-\beta)(1-\eta)]+(1-\beta)\eta(k+d)-\psi'(e)=0.$$

By setting $d=d^{cge}$ the above first-order condition coincides with ($A2$). □

Simulation

We suppose that the users' valuation is distributed according to the Continuous Uniform Distribution between $[-1,1]$ to obtain normalized platform's demand and overall users' valuation respectively equal to $D=1/2$ and $V=3/4$. We set $B=10$ and the per-user advertising revenues $a=16$.[38] We also set $H=1000$, $k=5$, $\psi(e)=5e^2$, $\beta=0.82$, and

[38] We consider that advertising revenues for a video that reaches 1,000 monetised views on YouTube are believed to amount to $18 which are split between the platform and the contributor (see, e.g., "How to become a YouTube millionaire" in the Financial Times, May 31, 2019). YouTube approximately keeps $8 and, with our assumed parameters, $aD=8$. We have set the benefit that the contributor obtains equal to his/her share of the advertising revenue, albeit we have offered a different interpretation in the model.

$\eta = 0.95$.[39] The left panel of Figure 1 illustrates the relationship between e^{SB} and e^{**} (when $d = 0$) as functions of γ. As γ increases, there is a higher chance that a stricter filter results in more frequent type-I errors. Accordingly, e^{SB} decreases with γ - a stricter filter is less desirable. Similarly, e^{**} decreases with γ as the platform anticipates that it will not admit original content more often. Note that for γ sufficiently high, the use of a filter would not be desirable. This is also because the right-holder can later detect infringing content and ask it to be removed. The optimal d is also a decreasing function of γ. In particular, for $\gamma = 0$, $d^* = 52.07$, whereas for $\gamma = 0.2$, $d^* = 41.76$. With the narrow view, d^{NV} is independent of γ, is higher and equal to 52.63. The relationship between d^*, d^{NV}, and γ is represented in the right panel of Figure 1.

In Table 2 we compare platform's, right-holder's, users', contributors' utilities, and overall welfare under the current policy and our policy prescription for the previous values of the parameters and $\gamma = 0.2$, so that $d^* = 41.76$. We find that overall welfare increases by 38% with respect to the current policy: while the right-holder is better off, the platform is hurt by the proposed policy that implements second best; moreover, users and contributors are worse off.

Proof of Remark 2

Suppose that the platform does not challenge a notice when $\gamma_R = 0$, that is (9) does not hold. Then, it will not challenge a notice for any $\gamma_R > 0$, either. Consider now the right-holder's incentive to reduce type-I errors knowing that the platform never challenges a notice. The right-holder's objective function would be: $-(1-\eta)Pr[\theta = 0|A]H - \phi(\gamma_R)$. As this is decreasing in γ_R, it is immediate to see that the right-holder does not have any incentive to reduce the occurrence of type-I errors. □

[39] Maintaining the assumption that the posted video would generate 1,000 monetised views, we assume that the right-holder's loss would amount to \$1 per view. Manually removing a video after receiving a take-down notice requires assigning dedicated employees (e.g., see the 2018 report "How Google Fights Piracy"), and we assume this to cost \$5, supposing that it takes 20 minutes to properly review a video and the moderator receives \$15 per hour, well below the median salaries at big US high-tech firms as these services are typically outsourced (e.g., see "The Trauma Floor" in the Verge, published on February 25, 2019). The justification for $\beta = 0.82$ comes from Kurdi et al. (2021), which shows that more than 17 per cent of videos posted on YouTube are deleted within a week. As some rule-infringing videos may be undetected, we round it up to 18 per cent. As for the value of η, automated recognition softwares used by right-holders are believed to detect a large fraction of infringing content.

Proof of Lemma 1

See that:

$$
\begin{aligned}
&\pi_R\Big(\gamma_R^I(e)\Big) \geq \pi_R\Big(\gamma_R^{II}(e)\Big)\\
\Leftrightarrow &-Pr[\theta=0|A](1-\eta)H - \phi(\gamma_R^I(e)) \geq\\
&-Pr[\theta=0|A](H-\eta d^P-\eta l^R) - Pr[\theta=1|A]\eta(1-\gamma_R^{II}(e))(d^R+l^R) - \phi(\gamma_R^{II}(e))\\
\Leftrightarrow &\eta\Big(Pr[\theta=1|A](1-\gamma_R^{II})(d^R+l^R) + Pr[\theta=0|A](H-d^P+l^R)\Big) \geq \phi(\gamma_R^I(e)) - \phi(\gamma_R^{II}(e)).
\end{aligned}
$$

□

Proof of Remark 3

It follows immediately from the text. □

Proof of Proposition 2

When the right-holder can make type-I errors, achieving second-best requires setting d^P and d^R so that the following system of implicit equations is satisfied:

$$
1-\frac{(1-\beta)(1-e^{SB})\eta(d^P+l^P)}{\beta(1-\gamma e^{SB})\eta[aD+d^R-l^P]} = \phi'^{-1}\left(\frac{\beta(1-\gamma e^{SB})\eta[aD+V+B+k]}{\beta(1-\gamma e^{SB})+(1-e^{SB})(1-\beta)}\right),
$$

$$
\eta\Big(Pr[\theta=1|A](1-\gamma_R^{II})(d^R+l^R)+Pr[\theta=0|A](H-d^P+l^R)\Big) \geq \phi(\gamma_R^I)-\phi(\gamma_R^{II}).
$$

The legislator should also choose d in such a way that (14) and (8) coincide. Since (13) holds for all e, given the policy (d^R, d^P, d^{**}), the platform cannot deviate by choosing a filter technology that induces litigation. Stated differently, the platform cannot choose a level of e lower than e^{SB} that prompts the right-holder to choose γ_R^{II} instead of γ_R^I, thereby inducing the platform to challenge the received notices. □

Proof of Remark 4

This is the condition that guarantees that the term in the square brackets in the left-hand side of equation (16) is strictly positive. □

Proof of Proposition 3

Note that e^{SB} is derived implicitly from the following first-order condition:

$$
\begin{aligned}
&\underbrace{[aD'(e^{SB})+V'(e^{SB})]}_{\geq 0}[A(e^{SB})-(1-e^{SB})\eta(1-\beta)]\\
&-[aD(e^{SB})+V(e^{SB})+B][\gamma\beta+(1-\eta)(1-\beta)]\\
&+\eta(1-\beta)k+(1-\eta)(1-\beta)H=\psi'(e^{SB}),
\end{aligned}
\tag{A4}
$$

where

$$D'(e) = v\frac{\partial\theta^E}{\partial e}f(-v\theta^E) \geq 0;$$

and

$$V'(e) = v\frac{\partial\theta^E}{\partial e}[1 + v\theta^E f(-v\theta^E)] \geq 0,$$

where the inequalities hold strictly when $v > 0$.

The platform would choose:

$$\max_{e\in[0,1]} \quad aD(e)[A(e) - (1-\beta)(1-e)\eta] - (1-\beta)(1-e)\eta(k+d) - \psi(e),$$

and the first-order condition yields:

$$\begin{aligned} &a\underbrace{D'(e^*)}_{\geq 0}[A(e^*) - (1-\beta)(1-e^*)\eta] \\ &-aD(e^*)[\gamma\beta + (1-\eta)(1-\beta)] + (1-\beta)\eta(k+d) = \psi'(e^*). \end{aligned}$$

Achieving second-best requires setting d in such a way that the platform internalizes the harm that copyright holders suffer when the infringement goes undetected as well as users' utility and the part of the contributor's private benefit that the platform does not extract. It is set so that the above equation coincides with (A4). □

Proof of Proposition 4

First, notice that the filter developer's payoff function is strictly concave in e for any T. Therefore, we can make use of the first-order approach and replace the incentive compatibility constraint by the first-order condition:

$$T = \frac{\psi'(e)}{(1-\beta) + \beta\gamma}.$$

After plugging the above transfer into the platform's objective function and taking the first order condition with respect to e, we obtain:

$$\frac{\partial}{\partial e}\left(e\frac{\partial\psi(\cdot)}{\partial e}\right) = (1-\beta)\eta(d+k) - aD[(1-\beta)(1-\eta) + \beta\gamma]. \tag{A5}$$

To achieve second best efficiency when the platform delegates the filter technology to a filter developer, the legislator should set d in such a way that (A5) coincides with (2):

$$d^{Del} = \frac{(1-\eta)H}{\eta} - \frac{(V+B)[\gamma\beta + (1-\beta)(1-\eta)]}{(1-\beta)\eta} + \frac{e^{SB}}{(1-\beta)\eta}\frac{\partial^2\psi(\cdot)}{\partial e^2}. \tag{A6}$$

It is immediate to see that $d^{Del} = d^* + \frac{e^{SB}}{(1-\beta)\eta}\frac{\partial^2\psi(\cdot)}{\partial e^2} > d^*$. □

Proof of Remark 5

The first-order conditions of the platform's problem yield:

$$\begin{aligned}-aD[(\overline{\gamma}-\gamma_1)\beta+(1-\beta)(1-\eta)]+(1-\beta)\eta(k+d)=&\psi_e(e,\gamma_1);\\ aD\beta e=&\psi_{\gamma_1}(e,\gamma_1).\end{aligned}$$

Note that e is continuously increasing in d. If d is set in such a way that $e(d)=e^{SB}$, $\gamma_1 << \gamma^{SB}$ because the platform does not internalize users' and contributors' benefits. As γ_1 is increasing in e, to reduce the gap between γ_1 and γ^{SB}, the legislator will induce $e(d)>e^{SB}$. □

Appendix B

Ex-ante versus ex-post incentives

In this appendix, we dig deeper into the reason behind copyright protection. Ex-post, when novel, original work has been created, granting an exclusive right to its usage and circulation to its creator involves a monopoly distortion. Thus, focusing only on an ex-post perspective, copyright protection might imply a social loss, as H might well be lower than $aD+V+B$. Yet, copyright protection is more solidly justified as a promised reward to encourage the development of innovative work. To better illustrate this point, we now extend the baseline model by assuming that some firm can first invest to develop some innovative work that, if protected by copyright, is worth H to this firm. When this firm succeeds in innovating, it becomes the right-holder of our model. The right-holder loses this amount H when some unoriginal content is uploaded by some contributor and kept on the online platform. Therefore, we now assume that, at stage 0, the firm (i.e., the possibly future right-holder) invests resources $x\in[0,1]$ at cost $\frac{cx^2}{2}$, with $c\geq H$. At stage 1, the game continues as in the baseline model, with the qualification that unoriginal content can be posted only if an innovative work has been created by the firm. As a result, from an ex-ante perspective, unoriginal work will be developed by the contributor with probability $x(1-\beta)$.[40]

Suppose first that there is no asymmetric information over the originality of the content that is developed by the contributor. If unoriginal content is kept online, in stage 0 the firm will choose:

$$x_1\in \arg\max_{x\in[0,1]} \quad x\beta H-\frac{cx^2}{2}\Leftrightarrow x_1=\frac{\beta H}{c}.$$

[40] We maintain the assumption that the probability that the content posted by the contributor is original is exogenous and independent of the copyright policy.

Expected welfare is:

$$EW_1 = x_1 H + \beta(aD + V + B) + x_1(1 - \beta)(aD + V + B - H) - \frac{c(x_1)^2}{2}.$$

If unoriginal content is taken out, $x_2 = \frac{H}{c}$ and expected welfare is:

$$EW_2 = x_2 H + \beta(aD + V + B) - \frac{c(x_2)^2}{2}.$$

By comparing the two expressions, we find that:

$$EW_1 > EW_2 \Leftrightarrow \frac{aD + V + B}{H} > \frac{1+\beta}{2\beta} \in [1, \infty).$$

If $H > aD + V + B$, it is never desirable to keep unoriginal content online: the ex-post harm suffered by the right-holder outweighs the benefits obtained by the other players when such content is kept online. However, even if $H < aD + V + B$, it may well be optimal to protect copyright so as to provide the firm with incentives to invest in generating innovative work. This is more likely to be the case when β is small. To see this, notice that the expression $\frac{1+\beta}{2\beta}$ is decreasing in β. Intuitively, the lower β the more depressed the firm's ex-ante incentives to invest when unoriginal content is allowed on the platform.

Let us now assume that there is asymmetric information over the originality of the content that is developed by the contributor. If unoriginal content is kept online, nothing changes with respect to EW_1. Conversely, if unoriginal content is taken out from the platform once detected, $x_3 = \frac{[\beta + (1-\beta)e + (1-\beta)(1-e)\eta]H}{c} > x_1$, whenever $e > 0$. Expected welfare is:

$$\begin{aligned} EW_3 =& x_3 H + \beta(aD + V + B) - x_3 \beta\gamma e(aD + V + B) \\ &+ x_3(1-\beta)(1-e)(1-\eta)(aD + V + B - H) - x_3(1-\beta)(1-e)\eta k - \frac{c(x_3)^2}{2}. \end{aligned}$$

Suppose for simplicity that $k = 0$. By comparing the two expressions, we find that:

$$\begin{aligned} EW_1 >& \quad EW_3 \Leftrightarrow \\ \frac{aD + V + B}{H} >& \quad \frac{(1-\beta)[e(1-\eta) + \eta][\beta(2-\eta) + e(1-\beta)(1-\eta) + \eta]}{2\{(1-\beta)\beta + [\beta + e(1-\beta)(1-\eta) + \eta(1-\beta)][\beta\gamma e - (1-\beta)(1-e)(1-\eta)]\}}. \end{aligned}$$

When β and γ are small, the legislator is more likely to protect copyright even when the ratio $\frac{aD+V+B}{H}$ is higher than 1. Likewise, a smaller k strengthens the case for copyright protection. Once again, copyright protection stimulates ex-ante investment. However, its benefit is reduced in the presence of asymmetric information.

References

Ali, N. S., Lewis, G., Vasserman, S., et al. (2022). Voluntary disclosure and personalized pricing. *Review of Economic Studies*, forthcoming.

Armstrong, M. (2006). Competition in two-sided markets. *The RAND journal of economics*, 37(3):668–691.

Beard, T. R., Ford, G. S., and Stern, M. (2018). Safe harbors and the evolution of online platform markets: an economic analysis. *Cardozo Arts & Ent. LJ*, 36:309.

Bone, R. G. (1997). Modeling frivolous suits. *University of Pennsylvania Law Review*, 145(3):519–605.

Buiten, M. C., de Streel, A., Peitz, and Martin (2020). Rethinking liability rules for online hosting platforms. *International Journal of Law and Information Technology*, 28(2):139–166.

Caillaud, B. and Jullien, B. (2003). Chicken & egg: Competition among intermediation service providers. *RAND journal of Economics*, pages 309–328.

Carroni, E. and Paolini, D. (2020). Business models for streaming platforms: Content acquisition, advertising and users. *Information Economics and Policy*, 52:100877.

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

Casner, B. and Teh, T.-H. (2023). Content-hosting platforms: discovery, membership, or both? *SSRN Working Paper.*

Cotter, T. F. (2006). Some observations on the law and economics of intermediaries. *Mich. St. L. Rev.*, page 67.

Crémer, J., de Montjoye, Y.-A., and Schweitzer, H. (2019). Competition policy for the digital era.

De Chiara, A., Ganuza, J. J., Gómez, F., Manna, E., and Segura, A. (2023). Platform liability with reputational sanctions. Technical report.

Elkin-Koren, N. and Fischman-Afori, O. (2017). Rulifying fair use. *Arizona Law Review*, 59:161.

Feher, A. (2023). How to enforce platforms' liability? Technical report.

Fromer, J. C. (2012). Expressive incentives in intellectual property. *Virginia Law Review*, pages 1745–1824.

Frosio, G. F. (2017). Reforming intermediary liability in the platform economy: A european digital single market strategy. *Nw. UL Rev. Online*, 112:18.

Gabison, G. A. and Buiten, M. C. (2019). Platform liability in copyright enforcement. *Colum. Sci. & Tech. L. Rev.*, 21:237.

García, K. (2020). Monetizing infringement. *UC Davis Law Review*, 54:265.

Grimmelmann, J. and Zhang, P. (2023). An economic model of intermediary liability. *Berkeley Technology Law Journal, Forthcoming.*

Hagiu, A. (2006). Pricing and commitment by two-sided platforms. *The RAND Journal of Economics*, 37(3):720–737.

Hornik, J. and Villa llera, C. (2017). An economic analysis of liability of hosting services: Uncertainty and incentives online. *Bruges European Economic Research Papers*, 37.

Hua, X. and Spier, K. E. (2021). Holding platforms liable. *Available at SSRN 3985066.*

Hua, X. and Spier, K. E. (2023). Platform safety: Strict liability versus negligence. *Available at SSRN.*

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

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

Jiménez Durán, R. (2021). The economics of content moderation: Theory and experimental evidence from hate speech on twitter. *Available at SSRN 4044098.*

Johnen, J. and Somogyi, R. (2021). Deceptive features on platforms. *CEPR Discussion Paper No. DP16175.*

Johnson, J., Rhodes, A., and Wildenbeest, M. R. (2020). Platform design when sellers use pricing algorithms.

Jullien, B. and Sand-Zantman, W. (2021). The economics of platforms: A theory guide for competition policy. *Information Economics and Policy*, 54:100880.

Kurdi, M., Albadi, N., and Mishra, S. (2021). “think before you upload”: an in-depth analysis of unavailable videos on youtube. *Social Network Analysis and Mining*, 11(1):1–21.

Landes, W. and Lichtman, D. (2003). Indirect liability for copyright infringement: Napster and beyond. *Journal of economic perspectives*, 17(2):113–124.

Landes, W. M., Posner, R. A., et al. (2003). *The economic structure of intellectual property law*. Harvard University Press.

Lefouili, Y. and Madio, L. (2022). The economics of platform liability. *European Journal of Law and Economics*, 53(3):319–351.

Liebowitz, S. J. (2018). Economic analysis of safe harbor provisions. *CISAC, February*, 27.

Madiega, T. A. (2020). Reform of the eu liability regime for online intermediaries: Background on the forthcoming digital services act.

Madio, L. and Quinn, M. (2023). Content moderation and advertising in social media platforms. *Available at SSRN 3551103*.

Menell, P. S. and Scotchmer, S. (2019). Economic models of innovation: stand-alone and cumulative creativity. In *Depoorter, B. And Menell, P. (eds): Research Handbook on the Economics of Intellectual Property Law*. Edward Elgar Publishing.

Polinsky, M. A. and Shavell, S. (1998). Punitive damages: An economic analysis. *Harvard Law Review*, 111:869–962.

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

Schruers, M. (2002). The history and economics of isp liability for third party content. *Virginia Law Review*, pages 205–264.

Spier, K. E. (2007). Chapter 4: Litigation. volume 1 of *Handbook of Law and Economics*, pages 259–342. Elsevier.

Teh, T.-H. (2021). Platform governance. *American Economic Journal: Microeconomics*, Forthcoming.

Urban, J. M., Karaganis, J., and Schofield, B. (2017a). Notice and takedown in everyday practice. *UC Berkeley Public Law Research Paper*, (2755628).

Urban, J. M., Karaganis, J., and Schofield, B. L. (2017b). Notice and takedown: Online service provider and rightsholder accounts of everyday practice. *J. Copyright Soc'y USA*, 64:371.

Zennyo, Y. (2023). Should platforms be held liable for defective third-party goods?

Zhang, P. (2021). The human side of cyber takedowns: Theory and evidence from github. *Available at SSRN 4274527*.

Tables and Figures

Table 1: Description of main variables and parameters.

β	Prob. that the content is original
e	Investment in the filter
$\psi(e)$	Costs of e
a	Per-unit advertising revenue
D	Users' demand
V	Overall users' utility
v_i	User i' utility from the content
H	Harm suffered by the copyright holder if the unoriginal content is kept online
η	Prob. that unoriginal content is detected by the right-holder
γ	Severity of platform's type I-errors problem
$1-\gamma^R$	Severity of right-holder's type I-errors problem
$\phi(\gamma^R)$	Right-holder's cost of reducing type-I errors
k	Platform's cost of processing take-down notices
B	Contributor's benefit from content posted online
d	Expected loss to the platform imposed by the legislator
δ	Fraction of advertising revenue that enters the welfare function
d^R	Damages paid by the right-holder
d^P	Damages paid by the platform
l^P	Legal expenses incurred by the platform
l^R	Legal expenses incurred by the right-holder

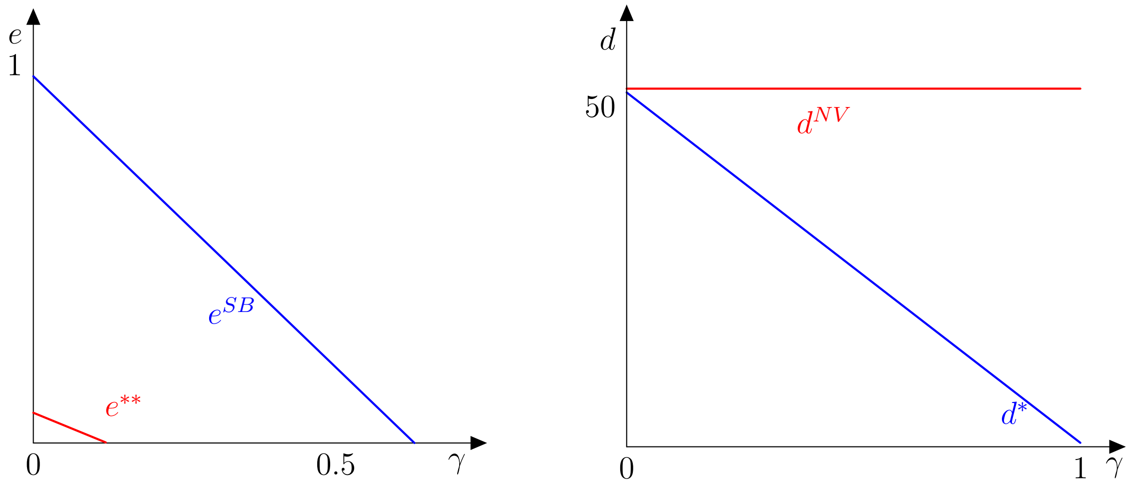


Figure 1: Filter technology and optimal d.

Table 2: Welfare comparison.In this simulation, we assume that $D = 0.5$, $V = 0.75$, $B = 10$, $a = 16$, $\beta = 0.82$, $\gamma = 0.2$, $\eta = 0.95$, $k = 5$, $H = 1,000$, and $\psi(e) = 5e^2$. In both the no liability scenario and under the current (default) policy, $e = 0$. Under the policy that implements second best, $d^* = 41.76$ and $e^* = e^{SB} = 0.66$. Under the narrow view, $d^{NV} = 52.63$ and $e^{NV} = 0.847$.

	Platform	Right-holder	Users	Contributors	Social Welfare
No liability	8	-180	0.75	10	-161.25
Default policy	5.7	-9	0.62	8.29	5.69
Implementing SB	0.82	-3.06	0.54	7.15	7.87
Narrow view	0.365	-1.38	0.52	6.82	7.7

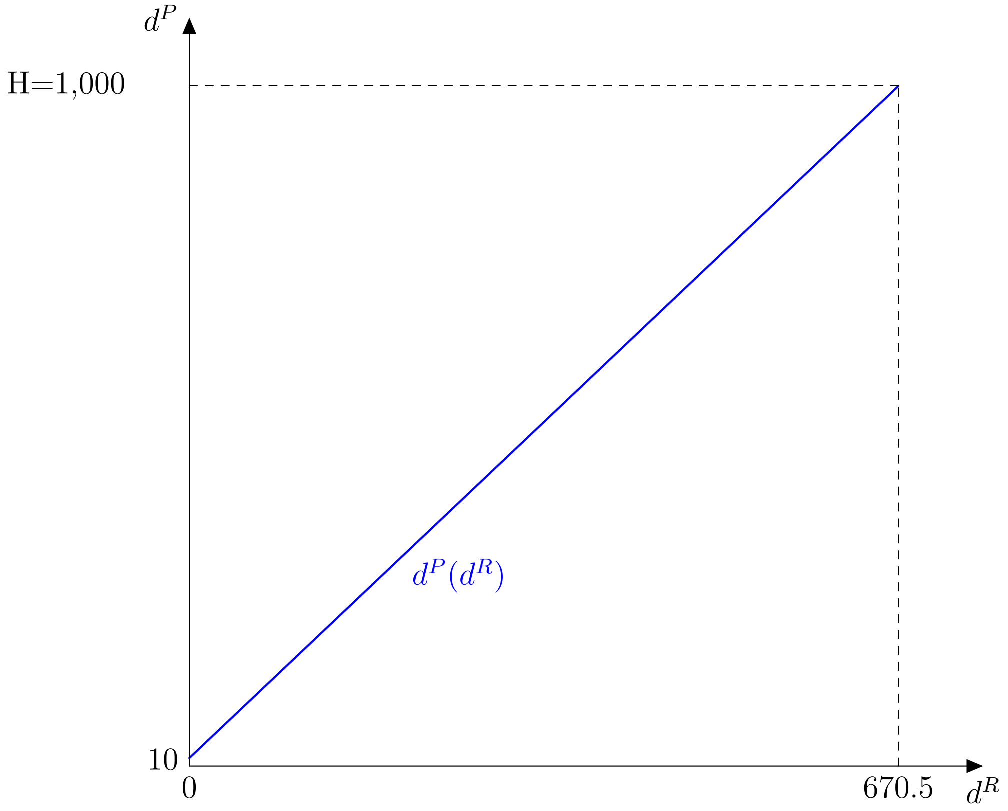
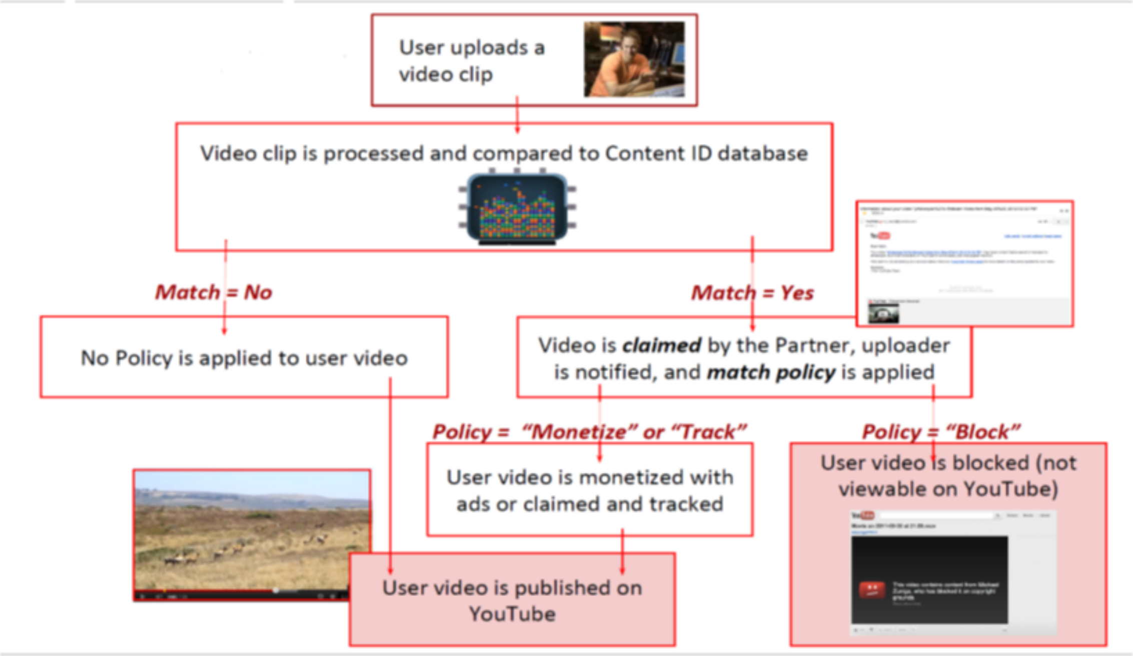


Figure 2: Relationship between d^P and d^R.

Minor infringements and content monetization - figure

Figure 3: YouTube's Content Id where a content flagged as infringing can be either blocked or monetized.


User uploads a video clip
Video clip is processed and compared to Content ID database
Match = No
Match = Yes
No Policy is applied to user video
Video is claimed by the Partner, uploader is notified, and match policy is applied
Policy = "Monetize" or "Track"
Policy = "Block"
User video is monetized with ads or claimed and tracked
User video is blocked (not viewable on YouTube)
User video is published on YouTube