International Review of Law & Economics 71 (2022) 106079

Contents lists available at ScienceDirect

International Review of Law & Economics

journal homepage: www.elsevier.com/locate/irle

Using experimental evidence to improve delegated enforcement[☆]

Lenka Fiala[a,*], Martin Husovec[b]

[a] Nova School of Business and Economics, Portugal
[b] London School of Economics and Political Science (LSE), United Kingdom

ARTICLE INFO

Article history:
Received 17 September 2021
Received in revised form 25 February 2022
Accepted 21 June 2022
Available online 25 June 2022

JEL classification:
C91
D02
K24
K42

Keywords:
Lab experiment
Notice and takedown
Online enforcement
Copyright
Content moderation
Alternative dispute resolution

ABSTRACT

Digital content today is governed by online providers like Facebook or YouTube. Increasingly, these providers are expected to enforce the law by removing illegal content, such as copyright infringement or hate speech. Typically, once they are notified of its existence, they have to assess it and, if infringing, remove it. Otherwise, they face liability. This system of content moderation is a form of delegation of the state's tasks to private parties. In literature, it is empirically established that some schemes of delegated enforcement can trigger substantial false positives, mostly due to over-compliance by providers and under-assertion of rights by affected content creators. This results in a phenomenon known as over-blocking: collateral removal of lawful content. We conduct a laboratory experiment to test a possible solution to this issue, as proposed by Husovec (2016). Our results show that an external dispute resolution mechanism subject to a particular fee structure can significantly reduce over-compliance by providers and improve the accuracy of their decisions, largely thanks to the content creators taking initiative. It does so by re-calibrating the typical asymmetry of incentives under the delegated enforcement schemes. The principles behind the solution have the potential to improve also other schemes of delegated enforcement where providers have weak incentives to properly execute delegated tasks in the public interest.

1. Introduction

Google alone blocked more than 5 billion links since 2011 on copyright grounds only.[1] In a single month, Twitter suspended 235 000 accounts for allegations of extremism.[2] Online providers are increasingly expected to act as agents of state by essentially doing the government's job – enforcing the law by removing illegal content. This delegation of responsibilities increases the efficiency of disputes but comes at a cost (Husovec, 2021). As recently pointed out by a member of the Court of Justice of the European Union, Advocate General Saugmandsgaard Øe.[3]

> Such a risk of an "over-blocking" exists, generally, where public authorities hold intermediary providers liable for illegal information provided by users of their services. In order to avoid any risk of liability, those intermediaries may tend to be over-zealous and excessively block such information where there is the slightest doubt as to its lawfulness.

Delegated enforcement tainted by over-blocking damages the digital business and online speech ecosystem. It exposes everyone who invests in digital presence to a risk of having it ruined in seconds. It also generally endangers the durability of the content placed in the digital space.

In the most typical scenario, if providers receive a notification about an alleged infringement, they have to act expeditiously to remove the content; otherwise, they can face liability of their own. In Germany in the area of hate speech, for instance, they can face fines up to 50 million EUR if they do not systematically remove some

[☆] Funding for the experiment provided by the Tilburg Law and Economics Center (TILEC) and the CentERLab is gratefully acknowledged. We also thank Eric van Damme and Jan Potters for their feedback, as well as participants of the Prague Conference on Behavioral Sciences (2018), and seminar participants at the London School of Economics and Political Science, University of Copenhagen, Lund University, Tilburg University, and Masaryk University.

[*] Corresponding author.
E-mail addresses: lenka.fiala@novasbe.pt (L. Fiala), m.husovec@lse.ac.uk (M. Husovec).

[1] transparencyreport.google.com/copyright/overview?hl=en

[2] nytimes.com/2016/08/19/technology/twitter-suspends-accounts-extremism.html

[3] Opinion of the Advocate General Saugmandsgaard Øe, Case C-401/19, Republic of Poland v European Parliament and Council of the European Union, ECLI:EU:C:2021:613, para 142.

https://doi.org/10.1016/j.irle.2022.106079
0144-8188/© 2022 The Author(s). Published by Elsevier Inc.
CC_BY_4.0

types of content within 24 h.[4] Under the EU Terrorist Content Regulation, they have to remove terrorist content within 1 h or face financial penalties of up to 4 per cent of their global turnover.[5] In copyright law, a failure to act upon notification exposes providers to damages and injunctions aiming to stop their services.[6]

This general choreography of privatized enforcement is widely known as *notice and takedown*. It allows a quick and cheap way of removing a large number of infringements from the Internet, while at the same time, it enables decentralized user-generated content to be shared without prior permission. It has been adopted more than 20 years ago when the user-generated content was slowly rising into prominence. However, the framework has long been suspected to have an over-blocking problem, at least in some areas of content or on some types of services.[7]

Simply put, over-blocking is often a rational choice *given* the existing legal framework. Faced with potential liability, providers not only strive to save resources on the assessment of notified content but also err on the side of caution. This is because providers as decision-makers face no punishments for wrongfully removing legitimate content. The liability systems often fail to create equally strong counter-incentives for providers to avoid over-removal by protecting content creators. Therefore, in the absence of natural *business* counter-incentives, everything points towards the removal of disputed content. The resulting *rational bias for over-blocking* in the enforcement chain is supported by the empirical literature in the area of copyright law (see Section 2).

The contribution of our paper is three-fold: First, we show how to model delegated enforcement in a laboratory experiment by operationalizing the above-described interactions and uncertainty about the true state of the world (i.e., legality of the content in question) faced by both players. Second, we provide experimental evidence that the compensated Alternative Dispute Resolution (ADR) as a solution increases the number of correct (final) decisions compared to the status quo without hurting the aggregate profits of the modeled players. Specifically, we find that our proposed system both nudges the providers not to use blanket take-down strategies, and it gives more power to the creators who can fight incorrect decisions by the providers. However, it is primarily the creator-initiated punishments and complaints that drive the overall decrease in the number of incorrect decisions. The providers (are able to) improve their accuracy only when evaluating a relatively simple case, suggesting that there is a natural upper bound of the effectiveness of this policy in the first stage, i.e., prior to creator complaints. And third, on a more general level, we demonstrate that restoring symmetry in the incentives of providers can improve the quality of decisions under delegated enforcement.

The paper is structured as follows: In Section 2, we review the existing literature on delegated enforcement, and establish three key features of the status quo: over-notification, over-compliance, and under-assertion. In Section 3, we present our model and hypotheses for our experiment, which is described in Section 4. Section 5 presents the results and elaborates on the limitations and practical implications of our findings. Section 6 concludes.

2. Literature

Given that most of the delegated enforcement takes place behind closed doors, empirical studies systematically reviewing the problems described above are difficult to organize. Inevitably, such studies of notice and takedown in practice have to rely on just a few available methods: (1) interviewing notifiers, providers and content creators (Urban et al., 2017); (2) experimental upload and subsequent notification of own or third party content (Nas, 2004; Dara, 2011; Perel and Elkin-Koren, 2017), (3) analysis of a few data sets shared publicly by providers, such as Lumen data[8] (Urban and Quilter, 2006; Urban et al., 2017; Seng, 2014, 2015), and (4) tracking of the public availability of the content over a pre-set period (Erickson and Kretschmer, 2018). So far, qualitative and quantitative studies were employed to understand the notification landscape and the assertion of rights by content creators. Then qualitative and experimental studies were used to observe over-compliance by providers. However, the existing evidence is mostly copyright-centered (for an exception, see Witt et al., 2019). In the area of hate speech law enforcement, the evidence is only anecdotal.

The empirical studies to date find the following:

Stage 1: *Over-notification*

In daily practice, we observe that economically motivated notice submitters, e.g., music rights holders and their authorized enforcement agents,[9] do not engage in sufficient quality control. They maximize their profit by sending as many notices as possible for as low a cost as possible. As a consequence, they *over-notify*. The reason for this outcome is mainly the fact that sanctions for over-notification (false positives) are rather limited, and thus notifying parties have little economic incentive to improve their quality control and reduce the resulting externalities they impose on others. Very often, the only real backlash is unwanted media attention. While the situation is slightly different in the context of notifiers who are not economically motivated (e.g., citizens notifying hate speech), the problem of the quality of notifications remains the same.

Urban and Quilter (2006) find that 31% of notifications in their sample had significant issues regarding the validity of the copyright takedown requests. Urban et al. (2017) worked with two different data sets. One concerning Google Search (Study 2) and the other concerning Google Image Search (Study 3). They found 28.4% and 36.8% of requests to be questionable on different legal grounds. Similarly, Seng (2015) finds that 8.3% of all takedown notices in his large sample failed to comply with the functional formalities. All these reported error rates constitute the most basic ‘procedural’ mistakes and do not include the more investment-intensive layer - the correct legal assessment of the content of such notifications. This shows that incentives for the quality of notifications which are the input of the notice and takedown choreography are not set right.

Stage 2: *Over-compliance*

After the submission of notifications, all the notices are processed by the providers who choose the extent and method of review. Increasingly, the providers use a lot of technological tools, such as content recognition or artificial intelligence. Theoretically, the providers could still completely limit the effects of over-notification by engaging in a thorough review of notices, thus taking down only infringing content. However, to evaluate each submitted notice, a provider has to first assess its legality and relevant facts, which is costly and often leads to uncertain outcomes. Moreover, no investment ever guarantees error-free decisions leading to a risk-free resolution. Furthermore, under-compliance can be punished by severe

[4] Netzwerkdurchsetzungsgesetz (NetzDG), available at https://www.gesetze-im-internet.de/netzdg/BJNR335210017.html.

[5] Article 3(3) and Article 18(3) of the Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online

[6] See the decision of the CJEU in YouTube/Cyando C-682/18 and C-683/18, ECLI:EU:C:2021:503.

[7] This paper does not have the ambition to establish the aggregate effects of over-blocking. It builds on the empirically substantiated assumption that at least in some areas (e.g., copyright enforcement) such effects are present and strong.

[8] The Lumen Deatabase, available at lumendatabase.org

[9] For the purposes of this paper, it is not important who exactly submits the notice. It may be regular users of the provider, copyright holders or their enforcement agents, or interest groups such as the Internet Watch Foundation.

fines, liability, or injunctions to stop service. In contrast, the legal risks of over-compliance are usually extremely low (for example, due to the existence of disclaimers, or litigation costs). More than legal implications towards their users (content creators), the providers worry about occasional unwanted media attention stemming from over-blocking, and its impact on the goodwill, especially if they market themselves as free speech champions (Klonick, 2017). As a consequence, they *over-remove*.

All the field experiments to date show that providers significantly over-comply with the false notifications received (Nas, 2004; Dara, 2011; Perel and Elkin-Koren, 2017). For instance, Nas (2004) created a website containing a public domain work of a famous Dutch author from 1871 and sent demands for its blocking by a fictitious right holder. From 10 Dutch providers, 7 removed or blocked the website containing copyright-free material, sometimes without notifying the website. This is despite the fact that the website clearly stated that the content is in the public domain and was written by Multatuli who is one of the most famous authors in the Netherlands.

The most extensive quantitative study to date is Urban et al. (2017). It provides evidence of the phenomenon of over-compliance on a large dataset of decisions of a wealthy incumbent. The authors show that while 70.3% of notices sent exhibited validity questions, Google removed 58.8% of the complained-of links. This shows that providers in practice do not sort through the low-quality notifications, either because they under-invest in screening, or prefer to take content down whenever there is any doubt about the content's legality. Furthermore, the study by Erickson and Kretschmer (2018) shows that takedown rates of parodies were not influenced by changes in their legality under copyright law but by what was dictated by right holders.

Finally, qualitative parts of the study Urban et al. (2017) largely confirm a tendency to remove if there is a doubt. Based on 29 interviews, providers report that the fear of liability might lead them to over-remove content.

It is worth emphasizing that where business counter-incentives against over-removal are sufficiently strong, the delegated enforcement can potentially avoid over-blocking. Such counter-incentives usually result from a particular subject matter that reduces the desirability of the service for consumers (e.g., child abuse material) or the type of service that is inherently linked with veracity or comprehensiveness. For instance, enforcement of the so-called right to be forgotten might be a potential service and type of content where providers have strong business incentives to mitigate over-blocking. Search engines, in particular Google, are sometimes required by law to delist upon request search results to digital content that is inaccurate, inadequate and irrelevant. In spite of the broad scope of the rules, they tend to dispute many requests, including before courts and authorities.[10] Therefore, one could argue that business counter-incentives are strong enough to minimize the over-blocking in this instance. However, even in this area, the complaints by affected websites whose links are to be delisted (e.g. news organizations) are instrumental in providing vital information to resolve the disputes. The strong interest to protect the quality of the product (veracity and comprehensiveness of search results) might explain why companies like Google invest more efforts in this area while having a less favorable track record in other areas, such as copyright law (Urban and Quilter, 2006; Urban et al., 2017; Seng, 2014, 2015).

Stage 3: *Under-assertion*

The affected authors of the content (e.g., YouTube or Facebook users who are content creators) usually do not take initiative to defend their content. They *under-assert* their interests. This can be due to intimidation, high legal risks,[11] and a weak prospect of a successful redress. Even in the areas of law where such legal redress – often dubbed counter-notice – explicitly exists today, it is massively underused (Urban and Quilter, 2006; Husovec, 2016; Bridy and Keller, 2015). Generally speaking, content creators often have either no or very weak rights to have their content reinstated after it was removed from the provider; moreover, providers can use re-design of their terms of service to circumvent any reinstatement.[12] In some cases, content creators are not even notified of their content being taken down.

The existing research shows that counter-notice, a complaint by affected content creators, is rarely filed in practice (Urban et al., 2017; Urban and Quilter, 2006; Seng, 2014). These findings are reinforced by a few transparency reports issued by companies that offer some additional (though limited) insights into the problem of under-assertion. According to available data, the counter-notice rate is frequently at rates below 1% of the *removed* content (Bridy and Keller, 2015; Klonick, 2020). As noted by Urban and Quilter, even in the United States, where an explicit complaint procedure exists, "the actual incentive to put back seems weak when compared to the incentives to take down" (Urban and Quilter, 2006).

Given the enormous amount of false positives observed even on the services owned by the biggest and wealthiest incumbents like Google (who presumably can invest the most in quality review), the underuse of counter-notice cannot be simply explained by a mere lack of interest of the affected parties in the blocked content. The rate of creator counter-notices disputing alleged infringements is often less than 1%, but the margin of error of notices is clearly higher according to the existing evidence (Urban and Quilter, 2006; Husovec, 2016; Bridy and Keller, 2015; Klonick, 2020).

The legal scholars and courts in Europe conceptualize the issue of over-blocking as a form of collateral censorship of legitimate speech (e.g. Witt et al., 2019; Kaye, 2019; Farrand, 2013; Randall, 2016; Tambini et al., 2007; Bar-Ziv and Elkin-Koren, 2018; Husovec, 2021). They have been long pondering on the arrangements of delegated enforcement because the out-sourcing of state responsibilities to private companies leads to challenges for legal accountability.[13] Among other things, legal scholars and policy makers are trying to find ways how to align decision-making by private companies with fundamental rights. Given that notice and take-down is a typical choreography of delegated enforcement, we hope to provide a generalizable example of how solutions in this area could be designed. Our message is that they should focus first on identifying incentives of providers and other players and then create meaningful counter-incentives to neutralize the rational bias to over-block legitimate content.

3. Theory

We model the notice and takedown enforcement system as a finitely repeated sequential game of two players: providers, and affected content creators. The reason we opt for a repeated game is to capture the inherent real-world aspect of the situation where content creators repeatedly interact with the same providers (e.g., as they build their Youtube channel).

[10] For a comprehensive review of Google's practices, see Theo Bertram and others, Five years of the Right to be Forgotten (2019), available at https://elie.net/static/files/five-years-of-the-right-to-be-forgotten/five-years-of-the-right-to-be-forgotten-paper.pdf

[11] By filing a complaint, the content creator exposes himself to further risk of liability towards the notifier. If the content was reinstated and found infringing, the content creator could additionally be held liable for damages along with the provider. Depending on the jurisdiction, such indemnification may be not negligible.

[12] Even when the law explicitly guarantees a remedy, its content is more about a possibility to be heard rather than a right to put the content back.

[13] The only comparable model in the literature Kim and Kim (2017) looks at the interaction of a petitioner and a search engine with a possibility to appeal to a higher authority.

At the beginning of the baseline game, the provider receives a notice. He has a strict time limit[14] to decide whether to keep the content up or take it down. This decision is communicated to the content creator, who, having had more time to familiarize himself with the notice,[15] chooses whether to submit a counter-notice (subject to a cost) or not.

If the creator chooses not to submit a counter-notice, the game ends. If he does submit it, a randomization device either implements this decision (25% probability) or not. This step aims to mimic how little attention is typically paid to the counter-notices by providers; we consider 25% to be a rather optimistic figure in this context. Notice, however, that the real-world lack of response to counter-notice is endogenous: by keeping it the same across treatments, we are limiting the in-game providers' actions, and prevent ourselves from studying this margin. Of course, a randomization device may be perceived very differently by the affected creators than a provider ignoring their counter-notice, even if the outcome is the same: for example, the randomizer explicitly communicates a level of risk that would be unknown otherwise, and it cannot be attributed (positive or negative) intentions. However, as long as creators in our experiment respond similarly to the device under all treatments, this does not compromise internal validity.

If the counter-notice is not implemented, the game ends. If it is implemented, the provider is informed about it and given the choice to reconsider his initial decision as from the provider's point of view, the content creators' complaints can provide new information or legal analysis. This reconsideration is final and ends the game as both players receive feedback about their own earnings.

To model the legal risk of keeping content online that should have been taken down, we penalize providers for this type of mistakes: After a given number of rounds, every such mistake leads to an additional ten percentage point risk for the provider to lose all their earnings from these rounds.[16] Content creators do not face this risk.[17]

While this interdependency between rounds for providers makes the game more difficult to analyze, we believe it is a worthwhile trade-off as it mimics the fact that in the real world, the provider learns from his mistakes with a delay, and a single bad decision can wipe out a substantial portion of profits (or shut down the business altogether).

The solution to this baseline over-compliance problem tested by this paper is then an external independent alternative dispute resolution (ADR) mechanism where a creator can direct his complaints subject to a fee. The ADR is a loosely defined informal way of resolving disputes that differs from typical arbitration or mediation solutions. Drawing an analogy with the domain name ADR systems, such as UDRP (Christie, 2002), it is very narrow in its application, immediately executable by a technical intermediary, and generally convincing by the power of specialized expertize. It is meant to provide for a quick resolution of narrow types of cases in an informal procedure. With the ADR, the creator gains a new way of challenging the decisions of the provider. After the complaint is filed and the creator pays a moderate fee, the ADR reviews the case and makes a decision. In the event that the ADR panel decides that content should be reinstated, the provider has to comply with it and fully compensate the creator's fee and pay additional fees to ADR. This should create an incentive for providers to further invest in the quality of their review in the long run.[18] For simplicity, we assume that the ADR body is always correct in its decision and hence we fully automate it.[19]

Table 1
Model Parametrization.

Variable	Value	Units
Starting endowment (both players)	10	tokens
Decision time provider	15	seconds
Time familiarization creator	30	seconds
Damage to creator from 'takedown'	4	tokens
Cost of counter-notice to creator	2	tokens
Damage from counter-notice to provider	1	tokens
Probability counter-notice is implemented	25	percent
Cost of ADR complaint to creator	5	tokens
Compensation to creator if won ADR ruling	8*	tokens

*Upon winning the ADR ruling, the damage done to the creator is automatically reversed, resulting in "additional" +4 tokens for the creator.

3.1. Parametrization and equilibria

We set the parameters of our game such that they capture the key real-life incentives in place. For an overview, see Table 1. Note that the game in the experiment is repeated in sets of five periods after which risk to providers from incorrect decisions to keep the content on their platforms is resolved; this is relevant for one of our equilibrium predictions as the other equilibrium predictions are stationary.

As explained in the literature section, a real-life counter-notice from a single user is unlikely to cause much damage to the provider, and if it does so, it usually tends to be in the form of media backlash. For this reason, we nullify this damage caused to the provider in case the provider revises his decision in line with the content creator's wishes, as in such a case there is much less of a reason for the backlash.

Notice that if the content creator wins the ADR ruling, the provider's decision is overturned: this means that the creator not only receives the compensation, but the damage done to him is reversed, mimicking the real-life situation that a creator can earn money from reinstated video content. (We abstract from the fact that some types of content, e.g., news coverage, may lose its revenue generating potential if not reinstated promptly.) Of course, in case the creator appeals a decision to keep the content online and wins the case, the content will be taken down and damage applied, offsetting the compensation the creator receives for winning the case and submitting the complaint.

[14] This time limit is meant to mimic the real-life time pressure put on providers to resolve cases quickly.

[15] We consider it a natural assumption that content creators are more familiar with their content than providers.

[16] In line with what is typically observed on online platforms, we mimic penalties for providers such as fines and injunctions to stop service.

[17] This reflects not the legal but practical situation of a usual lack of feasibility when trying to enforce the law against the content creators personally.

[18] This set-up fully replicates the UDRP model. There is no need to discuss the design of the decision-making within an ADR body, as our model does not try to design an optimal decision-making body. Admittedly, the institutional design would be important in practice but is beyond the scope of our paper. In fact, our solution can be also implemented by other bodies, such as state authorities, as long as the financing structure remains the same.

[19] Since the ADR is an independent expert body, the risk of false positives in their decisions should be very low. We expect the ADR in practice to filter litigation risk in a way similar to the UDRP, the alternative dispute resolution system for the domain names. See the work of Kur (2002), who finds that 5 out of 700 UDRP (0.7%) cases used in the sample were recorded as being pursued before the court after the decision of the UDRP panel, noting that "insofar as references to court proceedings following UDRP decisions have been published, they tend to confirm the general expectation that the decisions are rather seldom challenged in court".

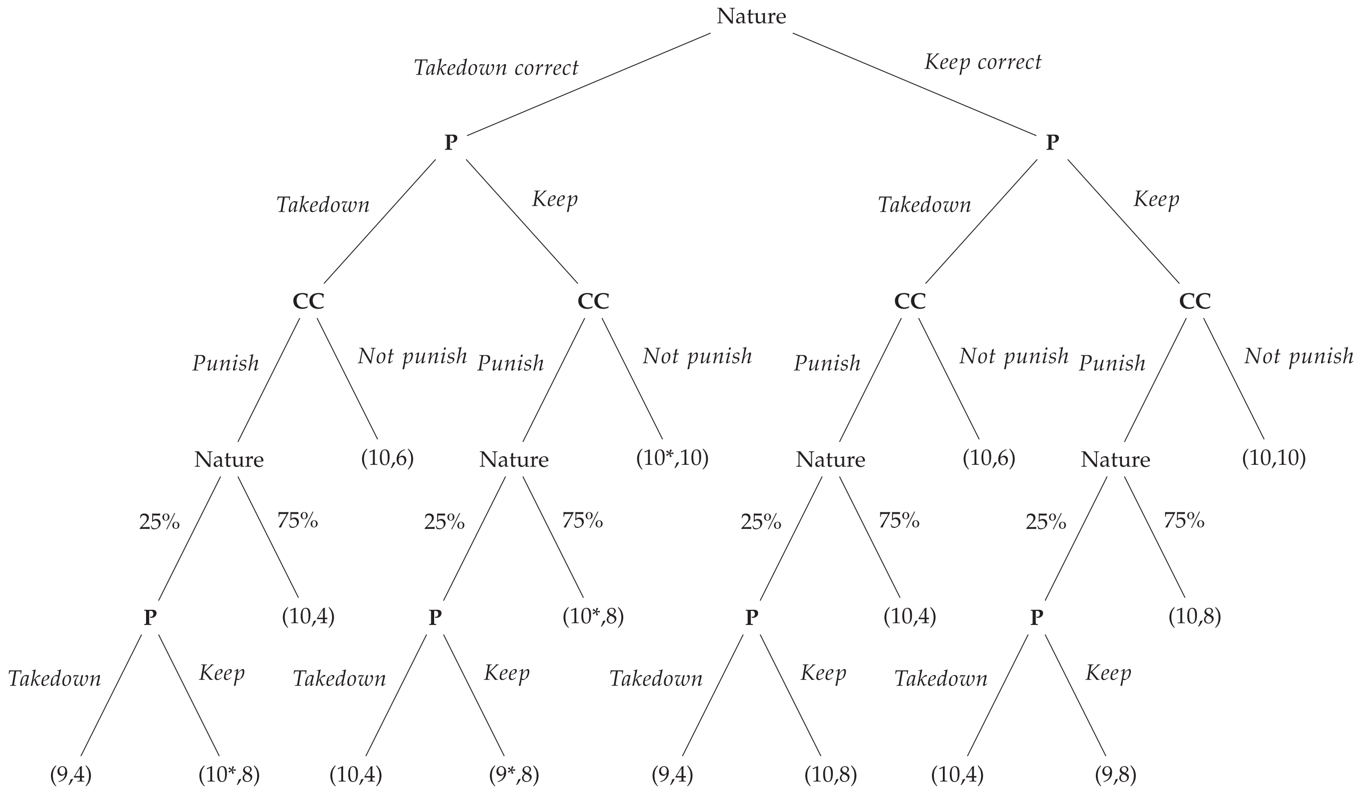

Fig. 1. Baseline Notice & Takedown Game. Decision nodes of providers (P) and content creators (CC) highlighted in bold. *denote incorrect decisions that expose the providers to the risk of losing earnings after a set of 5 periods. The first payoff in parentheses refers to the provider, the second payoff refers to the creator.

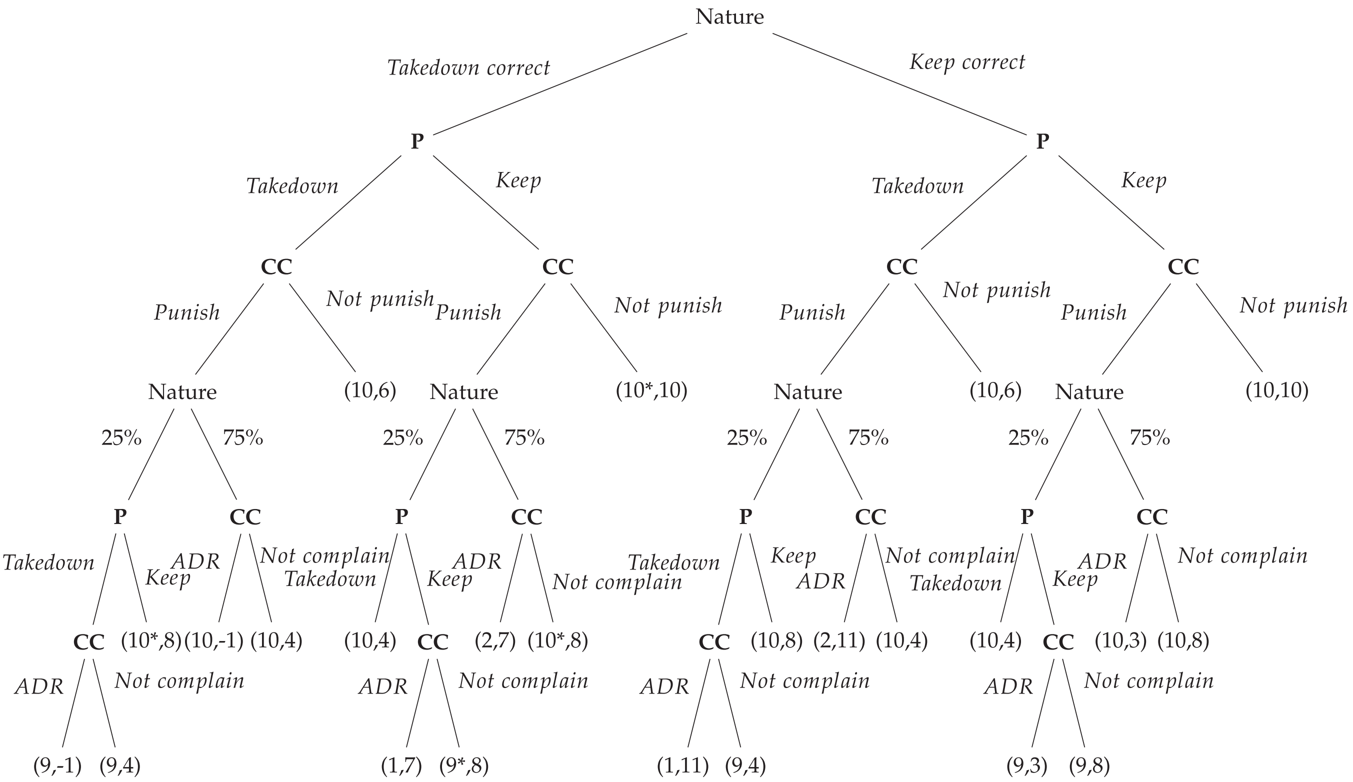

Fig. 2. ADR Extension of the Notice & Takedown Game. Decision nodes of providers (P) and content creators (CC) highlighted in bold. *denote incorrect decisions that expose the providers to the risk of losing earnings after a set of 5 periods. The first payoff in parentheses refers to the provider, the second payoff refers to the creator. Note that the ADR is assumed not to make mistakes.

For game trees representing this basic game structure and incentives, see Fig. 1 and Fig. 2. Note that * are used to denote objectively incorrect decisions that expose the providers to risk as discussed above.

We distinguish three information benchmarks that are appropriate in the real world context:

1. Full information: the true state of the world (which option, *takedown* or *keep*, is correct) is known by both players with certainty
2. Content creator informed: only the content creator knows the true state of the world with certainty; the provider considers both states equally likely
3. No information: neither player knows the true state of the world; both players consider both states equally likely

The first case corresponds to cases where law infringement is very easy to determine, such as a full-length movie video. The second case represents situations that are more difficult for the provider to judge, such as recognizing the difference between a parody and a copyright violation, which should be known to the content creator since he knows how the content was created. Finally, the third case models gray area situations, for example, those with insufficient precedent in that particular content domain. In that case, we consider it reasonable that both players operate with some uncertainty.

Of course, the model can be easily extended to accommodate any specific beliefs of the players; for simplicity of exposition, we only discuss the above three.

The analysis under perfect information is the simplest: using backward induction, it is easy to show that the ADR refines the baseline equilibrium outcomes. While under baseline it is possible to end up in a socially inefficient outcome under the *keep* state of the world, the existence of ADR allows the content creator to always seek justice, and be heard, resulting in the first-best outcome: the provider always correctly classifies content, and the content creator never has to appeal these decisions.

In the other two cases, players form an expectation about their payoffs, and we simplify the analysis by assuming that the uninformed party simply had too little time to study the content in question, and thus has not been able to update its prior, which we set at both states being equally likely.

In baseline, as soon as the provider is uncertain about the state of the world, he becomes “conservative” in his strategy and sticks with a safe *takedown*, resulting in a unique equilibrium outcome, a *takedown* followed by *no punishment*, regardless of the true state of the world.

With the ADR in place, we set up the game to make it profitable for the content creator to complain only when he is reasonably certain that he is right (to avoid frivolous lawsuits); for this reason, with both players uninformed the ADR does not improve upon the baseline outcome. When the content creator is informed about the true state, he will submit a counter-notice (and, if necessary, complain to the ADR), but only do so in case of over-compliance (so, the provider taking down content when it should have been kept up). For more details on the equilibrium outcomes see Table 2, and for their derivation see the Appendix.[20]

Notice that we allow the content creators to punish and appeal both *takedown* and *keep* decisions to alleviate concerns about experimenter demand. As shown later in the results, our subjects quickly learn not to punish/appeal decisions that benefit them, regardless of the true state of the world. We take this as evidence that the subjects understand the game's incentives, and they prefer maximizing their own earnings over trying to achieve “truth” or other non-material goals.

Table 2
Equilibrium Outcomes under Different Information Conditions.

	Baseline	ADR
Perfect information	Under the **takedown** state, the unique equilibrium outcome is *takedown* followed by *no punishment*. Under the **keep** state, there exist infinitely many equilibrium outcomes: the provider can choose any action, and the content creator always responds with *no punishment*.	The unique equilibrium outcome is the provider matching the state of the world with his action (playing *takedown* in the **takedown** state, and playing *keep* in the **keep** state), to which the content creator responds with *no punishment*.
Only CC informed	The unique equilibrium outcome is *takedown* followed by *no punishment*, regardless of the state of the world.	The equilibrium of this game has a dynamic aspect: over the course of five periods, the provider plays *keep* exactly four times, to which the creator responds with *no punishment*, and the provider plays *takedown* once, which is *punished* (and, if needed, *challenged at ADR*) only if the true state of the world is **keep**. All orderings of four *keep* and one *takedown* strategies constitute an equilibrium.
No information	The unique equilibrium outcome is *takedown* followed by *no punishment*, regardless of the state of the world.	The unique equilibrium outcome is *takedown* followed by *no punishment*, regardless of the state of the world.

3.2. Hypotheses

In line with the above theoretical discussion, we hypothesize the following:

H1a: In the baseline condition, the providers exhibit an overblocking bias, disproportionately taking down more content than would be legally required.

H1b: In the baseline condition, the content creators under-assert their rights, i.e., do not submit counter-notices whenever providers make incorrect decisions.

H1c: In the baseline condition, counter-notices that are submitted will not change the providers' decisions.

H2a: In the ADR condition, the providers exhibit lower overblocking bias than under the baseline condition.

H2b: In the ADR condition, the accuracy of provider decisions improves compared to the baseline condition.

H3: The extent to which the ADR improves upon baseline in terms of correct decisions depends on the decision difficulty: The ADR has the lowest effect in situations that are difficult for both players to evaluate.

We do not formulate specific hypotheses about the relative importance of behavioral responses of the two players under the ADR; we are however interested in whether both of the players respond to changed incentives.

4. Experimental design

We implement the above described finitely repeated sequential game in a context-free environment, i.e., not linking it to the real-world provider-content creator interaction.

In our game, providers and content creators are referred to as players A and B, respectively. The provider makes a yes/no decision

[20] It has to be noted, however, that in all of these cases we assume standard, risk-neutral economic preferences, where agents are only concerned with their own monetary payoff. If we allowed for social preferences or risk aversion, we could obtain equilibria with more punishment, for example.

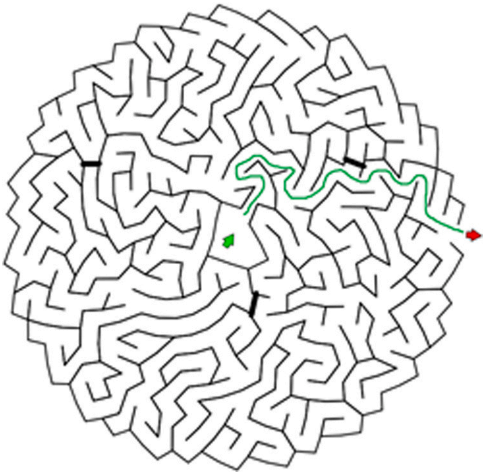

Fig. 3. An example of a maze with a solution.

under time pressure whether a solution to a maze exists.[21] (By a solution we mean a person can walk from one end to the other, see Fig. 3). This represents the decision to either take content down (yes) or keep it (no) after receiving the notice. At the same time, the content creator is given more time to study the maze (as a content creator would likely be familiar with his own content), but his evaluation of the case is not asked.

The reason we introduce time pressure is to mimic the real-life aspect where providers are required to make content control decisions expeditiously. Due to random assignment to treatment, we have no reason to believe that the subjects' underlying maze solving skills under time pressure affect our results.

Players receive equally many 'yes' and 'no' mazes, to match our theoretical set-up that operates with a prior of both answers being equally likely.

To mimic the various possible informational settings, we use puzzles of different difficulty levels.[22] For purposes of our analysis, we assume that smaller mazes, i.e., those classified as suitable for younger children, are easier than larger ones. For robustness checks, see the Appendix.

Counter-notice is in the experiment called *punishment*, and the ADR is called *a formal resolution body* to which a creator can *complain* if punishment previously failed to change the provider's decision. All other aspects of the game (structure of interactions, payoffs) were kept the same as in the above game description.

Notice the provider is immunized from any legal risk once he complies with the ADR decision. In line with our theoretical framework, the ADR is programmed not to make any mistakes. For details and a subject comprehension quiz, please see the instructions in the Appendix.

4.1. Procedures

We conducted our experiment at the CentER Lab, Tilburg University, in spring 2018 and 2019. Our 148 subjects (57% female), split 72:76 between baseline and ADR treatments, were students of social sciences. A typical session lasted 90 min and subjects earned 18 Euro on average. The show-up fee was 4 Euro, and the exchange rate between lab currency and Euro was 5 experimental dollars = 1 Euro. Software zTree (Fischbacher, 2007) was used to run this experiment.

Each session consisted of two parts: first, in both treatments, the subjects played three sequences (sets) of 5 decisions under the baseline specification. This was followed by a second part with three sets of 5 decisions, either in the baseline or ADR specification, depending on the condition. Unless specified otherwise, only the second half of the experiment (last fifteen periods) from each session is used for analysis to allow for sufficient learning of subjects. One set of decisions from each part was randomly selected for payment at the end of the experiment.

Subjects switched partners only between the two parts but switched roles after each 5-decision sequence to allow them to learn the incentives of the game faster. We created matching groups of 4 to maximize the number of independent observations.[23] We have 35 independent observations in total, split equally between treatment and control.[24]

We provide two layers of analysis, using the between-subject aspect of our design to compare baseline and the ADR treatment, and the within-subject aspect to make inferences about how maze difficulty, i.e., information, affect the outcomes.

5. Results

To see whether the ADR could be a feasible solution to our problem, we first need to show that our baseline condition sufficiently well mimics the real-world situation, i.e., exhibits the problem of over-compliance (H1a), under-assertion (H1b), and counter-notice ineffectiveness (H1c). This is done in the first subsection. Next, we show how the ADR changes subjects' behavior, and which consequences it has for the players (H2). In subsection three we explore which equilibrium outcomes and which informational settings these results are consistent with (H3). We discuss the implications of our results in subsection four.

Throughout, we use $\alpha = 0.05$ as our significance threshold and rely on non-parametric tests due to the non-normality of the data. Our tests and tables use aggregated data on the level of independent observations (4-person groups) unless indicated otherwise.

5.1. Baseline

As we argued in the Background subsection, the status quo is characterized by three main effects: .

1. Providers make systematic errors: more so in the dimension of over-compliance (taking down too much legitimate content)
2. Content creators typically do not fight back when such an error is made (under-assertion)
3. Even if content creators fight back, they usually remain powerless (and their content is kept down)

Looking at our experiment, and using only the second part of the experiment (i.e., last 15 rounds) to make sure both players are sufficiently experienced and understand the game, we argue that by and large, these three conditions were satisfied in our baseline treatment, and thus we have convincing support for our hypotheses H1a-H1c.

[21] We adjusted mazes from krazydad.com/mazes/.

[22] The source of our puzzles, krazydad.com/mazes/, separates mazes into 5 difficulty levels; for our experiment, we use the three easiest categories. We always mix 6 "easy" and 9 "difficult" puzzles for each part of the experiment.

[23] In one of our ADR sessions the subjects played four baseline sequences and only two ADR sequences due to a computer error. This resulted in two independent observations consisting of 8 subjects rather than 4. We provide robustness checks in the Appendix to show that this session does not drive our results.

[24] Unless indicated otherwise, an independent observation refers to the average behavior of subjects in their matching group over the last 15 rounds of the experiment.

Table 3
Baseline: Answers by Providers.

	True state of the world	
	Takedown	Keep
Answer *takedown*	35%	**30%**
Answer *keep*	**12%**	23%

The table provides an overview of the percentage of *takedown* and *keep* answers (out of total answers) depending on the actual correct answer (state of the world). Average group decisions are reported. Mistakes are highlighted in **bold**. Results are almost identical if revised answers by the provider are taken into account.

Table 4
Baseline: Learning to Err (Providers).

	First 15 Rounds	Last 15 Rounds
Mistakes as % of answers	28%	42%
...of which % *takedown* mistakes	59%	72%

The table provides a comparison of the *takedown*-direction mistakes in the baseline treatment. Average decisions of groups are reported. Results are almost identical if revised answers by the provider are taken into account.

Table 5
Baseline: Learning Not to Counter-notice (Content Creators).

	First 15 Rounds	Last 15 Rounds
Total # counter-notices	65	44
# Counter-notices to *keep* decisions	3	1
...as % relative to all *keep* decisions	(1%)	(0.5%)
# Counter-notices to *takedown* decisions	62	43
...as % relative to all *takedown* decisions	(20%)	(12%)

The table provides a comparison of how content creators respond to different provider decisions in the baseline treatment. Group decisions are reported.

Over the last 15 decisions the providers had to make, they made 42% incorrect decisions at the beginning of the game (i.e., prior to the creators' responses). Of these, the vast majority (72%) were in the over-compliance direction (see Table 3 and Table 4).

The fact that these mistakes are favouring *takedown* is no accident, and, in fact, a result of learning. Doing a within-subject comparison (so, comparing the first vs. the last 15 rounds of baseline), we see a greater number and greater proportion of over-compliance mistakes in the experienced sample (see Table 4).

The increase both in the number of overall mistakes, and the proportion of *takedown* errors as shares of total errors are significant on the 5% level (Wilcoxon sign-rank test p-values of 0.0002 and 0.0285; corresponding test z-statistics − 3.67 and − 2.18; N = 18 independent observations).

We, therefore, conclude that in our baseline treatment, the providers learn to err on the side of caution, i.e., *takedown*, providing support for our Hypothesis 1a.

This relatively large number of *takedown* errors allows us to verify our second criterion for our baseline to reflect reality: namely, do content creators fight back these types of mistakes, i.e., submit counter-notices?

As Table 5 makes clear, our subjects do learn not to fight: *keep* decisions are generally not subject to counter-notice, whereas *takedown* decisions are primarily targeted in the early stages, but the content creators learn not to do so (since it is ineffective, as will be shown below).

We can show there is significantly less overall counter-notice in the last 15 rounds (Wilcoxon sign-rank test p-value of 0.0215; corresponding test z-statistic 2.32; N = 18 independent observations).[25] This allows us to conclude that indeed, our content creators exhibit under-assertion, providing support to our Hypothesis 1b.

Table 6
ADR: Answers by Providers.

	True state of the world	
	Takedown	Keep
Answer *takedown*	29%	**19**%
Answer *keep*	**18**%	33%

The table provides an overview of the percentage of *takedown* and *keep* answers depending on the actual correct answer. Average decisions of groups are reported. Mistakes are highlighted in **bold**. Revised answers of providers are considered separately.

Our third criterion dictates that for our baseline to mimic reality well, the providers should not change their minds in response to counter-notice. We do observe that providers switching their answers is a rare occurrence, however, due to a small number of observations, we cannot claim a significant change in providers' behavior (i.e., learning) from the first 15 to the last 15 periods. In total, we only observe 4 switches in the first 15 periods and 1 switch in the last 15 periods.

While we cannot claim that our subjects learned not to switch in response to a counter-notice, we would like to underscore that the switching rate was low to begin with, and with a single switch observed among the experienced subjects, we are confident our setting provides suggestive evidence for Hypothesis 1c.

Taken together, we believe we set up a system in which providers disproportionately take down too much legitimate content, content creators do not often fight these decisions (but fight them often enough for us to analyze), and when they do so, the content creators are basically powerless: the providers' decisions do not change. This constitutes our benchmark, closely replicating the real-world status quo, that we pit our treatment of ADR against.

5.2. ADR

Next, our aim is to test whether the ADR improves the status quo; namely, whether it decreases over-blocking.

First, the over-compliance bias is indeed mitigated: now both over-compliance and under-compliance are almost equally common (see Table 6).

As a between-subject analysis shows, there are significantly fewer mistakes in the direction of over-compliance as a fraction of total mistakes under ADR compared to the baseline (Wilcoxon rank-sum test p-value of 0.0049; corresponding test statistic − 2.76; N = 35 independent observations). This lends support to Hypothesis 2a.

However, a valid criticism of our mechanism could be that it merely incentivizes the providers to err in the opposite direction (i.e., undercomply). To explore whether the ADR mechanism leads to a decrease in the overall error rates, we have three redress channels to consider: (i) the initial decision of providers, (ii) the response of providers to counter-notice, and (iii) the ADR itself imposing the correct decision if requested to act. We analyze these in turn.

In the first stage, i.e., the initial provider decision stage, the total number of mistakes (of either kind) decreases under ADR but this decrease (in this first stage of decisions) is insignificant (Wilcoxon rank-sum test p-value of 0.1527; corresponding test statistic − 1.44; N = 35 independent observations).[26] We therefore cannot conclude

[25] Due to a small number of observations we do not perform subsample tests for *keep* and *takedown* responses separately.

[26] In absolute terms, the number of *takedown* errors significantly decreases (p-val 0.0059; corresponding test statistic − 2.71; N = 35 independent observations) while the *keep* errors insignificantly increase (p-val 0.0585; corresponding test statistic 1.89; N = 35 independent observations).

Table 7
ADR: Counter-Notice Behavior (Content Creators).

	Baseline	ADR
Total # Counter-notices	44	114
# Counter-notices to Keep decisions	1	4
…as % relative to all Keep decisions	(0.5%)	(1.5%)
# Counter-notices to Takedown decisions	43	110
…as % relative to all Takedown decisions	(12%)	(44%)

The table provides a comparison of how content creators respond to different provider decisions in the (last 15 rounds of) baseline and the ADR treatment. For ease of exposition, group decisions are reported, rather than average decisions of independent observations.

Table 8
ADR: Switching after Punishment (Providers).

	Baseline	ADR
Total # implemented punishments	7	30
Total # switches	1 (14%)	17 (57%)

The table provides a comparison of how providers respond to being punished in the (last 15 rounds of) baseline and ADR treatment. Individual decisions are reported, rather than average decisions of independent observations (groups of four).

that our system actually improves accuracy as opposed to swapping one type of mistake for another in the first stage of decisions.

Second, the content creators under the ADR system do fight incorrect decisions of providers at much higher rates than in the baseline (Wilcoxon rank-sum test p-value of < 0.0001; corresponding test statistic 3.99; N = 35 independent observations), and do so overwhelmingly in cases when the provider chose *takedown*, as full 44% of *takedowns* are met with a counter-notice (see Table 7).

Of course, the mere presence of counter-notice is merely a possibility for redress, and hence we look at whether the content creators' complaints actually succeed in changing outcomes.

Looking at revised answers after counter-notice (but prior to potential ADR ruling), we see that this time many more providers switch than in the baseline (see Table 8). However, as shown by the Wilcoxon rank-sum test, this increase is not significant, likely due to the small number of observed switches in general (p-value of 0.0650; corresponding test statistic 2.02; usable N = 20 independent observations).

Finally, we check whether these switches improve accuracy relative to baseline. We, therefore, compare these second-stage answers to the baseline final answers. While the error rate does decrease from 42% to 36%, this is not significant (Wilcoxon rank-sum test p-value of 0.0803; corresponding test statistic − 1.76; N = 35 independent observations). Therefore, if the ADR significantly improves outcomes, this has to be driven by the third channel: the ADR decision body itself.

We thus turn to the ADR use by the content creators. Since we observed a total of 114 punishments, of which only 30 were implemented, and only 17 led to a decision change, this leaves us with a total of 84 + 13 situations where the ADR could have been pursued. Since 90 out of these 97 opportunities were indeed taken by the content creators, we see this as important evidence that the ADR is not used as a mere threat by content creators, but rather as a tool to resolve a dispute.

As a between-subject comparison reveals, while 58% of cases are in the end decided correctly (i.e., in line with the true state of the world) under baseline, this number jumps to 75% under the ADR, which is a significant increase (Wilcoxon rank-sum test p-value of < 0.0000; corresponding test statistic 4.29; N = 35 independent observations).

Table 9
ADR vs. Baseline: Change in Profits.

	Baseline	ADR
Profit providers	9.44	8.22
Profit content creators	7.24	8.19
Total profits	16.68	16.41

The table provides a comparison of average per-period profits for providers in the (last 15 rounds of) baseline and ADR treatment. Average outcomes of independent observations (groups of four) are shown. Profits are reported as amounts of tokens; both players started each round with an endowment of 10.

Table 10
Welfare Analysis.

	Baseline	ADR
Fewest incorrect (initial) decisions		(n.s.)
Fewest incorrect (final) outcomes		✓
Profits of providers	✓	
Profits of content creators		✓
Total profits	(n.s.)	

A ✓denotes the condition that performs better on the given criterion. In case there are no statistically significant differences between conditions, *n.s.* is used.

Taking all three of our main points together, we can conclude that the ADR indeed mitigates the problem of over-enforcement: the providers change their behavior both in their initial and revised judgments in favor of the content creators (in support of Hypothesis 2a), the ADR is used as the last resort to resolve disputes, and its presence significantly increases the number of correctly evaluated cases (in support of Hypothesis 2b). This increase primarily occurs in the final stage of the game, i.e., when the ADR is asked to evaluate the case. Note that we assume the content creators can always complain to the ADR, and the ADR is designed to impose the correct decision with 100% certainty.

Since this improvement in final outcomes comes at a cost (as both punishment and ADR complaints are expensive), it is interesting to look at how profits change under the ADR: As shown in Table 9, average per-period profits of providers decrease (Wilcoxon ranksum test p-value of 0.0005; corresponding test statistic − 3.35; N = 35 independent observations) and of content creators increase (Wilcoxon ranksum test p-value of 0.0006; corresponding test statistic 3.30; N = 35 independent observations). Overall, the total average profits remain unchanged (p-value of 0.2096; corresponding test statistic − 1.27; N = 35 independent observations).

Overall, whether the ADR improves upon the status quo requires a value judgment: Depending on the relative importance of the outcomes of providers, content creators, and the objective correctness of judgments, which system performs "better" can change. In Table 10 we provide a summary of the most basic evaluation criteria for comparison to illustrate this point.

Of course, this welfare analysis is incomplete; our game does not take into account other important players in the real world: notifiers, rights-holders (relevant for copyright infringement disputes), regular users of platforms, and the society at large. We consider this a fruitful avenue for future research.

5.3. Relation to theory

While the previous analysis demonstrates that the ADR in the aggregate changes outcomes in comparison to baseline, and the behavioral patterns are the closest to the perfect information benchmark (with providers sometimes making mistakes and to some extent redressing them under the ADR), we are interested

Table 11
Easy vs. Difficult Puzzles.

	Baseline		ADR	
	Easy	Difficult	Easy	Difficult
Decisions correct	61%	56%	74%	55%
Incorrect *takedown* punished	27%	14%	79%	61%
Incorrect *takedown* ADR complaint	–	–	96%	94%

The table provides a comparison of the player's behavior in the (last 15 rounds of) baseline and ADR treatment. The table shows the percentage of providers whose initial decision was correct, the percentage of content creators who punished an unjust *takedown*, and the percentage of content creators who complained to the ADR when facing an incorrect *takedown*.

whether this result depends on the difficulty of the case to be evaluated.

We exploit the fact that we have mazes of differing difficulty levels to see whether it is the easy puzzles driving our results, or whether the players behave similarly regardless of the maze difficulty. Specifically, we are interested in whether the use of ADR differs depending on the puzzle difficulty (our Hypothesis 3):

Just like in our previous analysis, unless specified otherwise, we look at only the last 15 periods to make sure we study the decisions of subjects who had ample time to learn the game.

First, as shown in Table 11, we see that the mere threat of ADR improves the providers' accuracy at the point of their initial decision, but only for easy puzzles (Wilcoxon ranksum test p-value of 0.0117; corresponding test statistic 2.50; N = 35 independent observations).[27] This suggests that the difficult puzzles are too difficult for providers to solve under time pressure, and even with the greater effort they cannot improve the accuracy of their decisions. We also see much higher counter-notice rates[28] under the ADR (as expected), and very high ADR complaint rates, suggesting that our subjects understand that threatening the use of ADR by costly counter-notice and not using it is not a profitable strategy. Importantly, we see that while fewer content creators submit a counter-notice when facing a difficult maze, they still pursue this strategy relatively often, suggesting that the content creators have at least some idea about the correct solution (state of the world).

Second, zooming in on the last column of Table 11, we see that generally when providers make *takedown* mistakes, they are punished and taken to the ADR, albeit somewhat less often in the difficult cases; the question remains as to what extent the providers seem to anticipate this. They could expect a lot fewer punishments/ADR complaints for difficult mazes if they think that the content creators (also) do not know the correct answer.[29]

However, as Table 12 makes clear, this is not the case: Regardless of the type of the puzzles, the providers behave as if they expected the content creators to punish them if they over-comply. This result however could be explained by other factors than purely information availability, such as social preferences that are outside our model.

Taken together, we can conclude that throughout our game, the players behaved similarly regardless of the puzzle difficulty they faced with the exception of providers in the first stage, as they improve their accuracy for easy puzzles under the ADR. Importantly, the providers err on the side of over-compliance just as often in both easy and difficult situations. Also, both types of puzzles drew counter-notice and ADR punishment, providing only mixed support for Hypothesis 3. This behavior is not consistent with our no-information theoretical benchmark.

Table 12
Difficult Puzzles: Behavior of Providers.

	ADR	
	Easy	Difficult
Initial *takedown* decisions	48%	49%
...of which incorrect	39%	40%

The table compares the initial decisions taken by the providers in the (last 15 rounds of) the ADR treatment for easy and difficult mazes.

5.4. Discussion

Our results show that we were able to successfully recreate the notice and takedown dynamics in the laboratory experiment. We create a situation where providers over-comply with *takedown* requests and disproportionately take down content that is legitimate, thereby overwhelmingly erring on the side of caution.

We also recreate the content creator's apathy regarding complaints. Not surprisingly, they complain less the longer they play the game. This shows that previous experience with a lack of credible remedy to their situation makes them even more resigned. Again, it should be underscored that our success rate for punishment is still much more optimistic than most of the real-world scenarios, where a 25 per cent chance of success of causing even small harm to the provider is unlikely.

The number of complaints markedly increases with the ADR option. In fact, almost all cases in which a content creator decided to complain are also followed up by an ADR filing. This, in fact, leads to a key improvement in the system's accuracy since the complaints are successful (because they are legitimate) in 64% (58 out of 90) of the cases.[30]

Importantly, looking at final outcomes, the introduction of the ADR substantially improves the system accuracy: The percentage of correctly evaluated cases increases from 58% to 75%. This improvement comes at a cost to providers, whose profits are redistributed towards the content creators; however, the aggregate profits remain the same.

The key channel for the improvement in accuracy is the ADR itself: while in the easy cases, its presence is enough to improve the provider's accuracy and mitigate the over-blocking bias, the key improvements are realized only when the ADR is actually used by the content creators. While potentially expensive, this channel has the major advantage over others that it (by assumption) cannot lead to under-enforcement.

Of course, under-enforcement remains a plausible outcome in theory: suppose that the providers were punished more severely for their mistakes if taken to the ADR, or the punishments were at least *perceived* as more severe: in that case, in fear of ADR, the providers might try to pre-empt creator complaints, and keep notified content online, causing under-compliance as a result. For this reason, we recommend careful calibration of ADR penalties in practice.

It is noteworthy that these effects are observed with a relatively high fee compared to the value that is taken away by the wrongful takedown (5 fee, 4 value). Moreover, the payment is framed as a fee

[27] Since we had fewer easy than difficult puzzles, this improvement was not enough to improve the system's overall accuracy already at the first decision stage, see the discussion between Table 6 and Table 7.

[28] The counter-notice rates are significantly higher both for easy and difficult puzzles, aggregate Wilcoxon ranksum text p-value of 0.0001; corresponding test statistic 3.71; usable N = 34 independent observations.

[29] Since the baseline benchmarks are not helpful in terms of disentangling which state we are in, we only explore the ADR treatment in detail.

[30] Somewhat ironically, we observe four cases where the content creator correctly complaints that the provider is wrong, but does so when the incorrect answer is to *keep* the content. We see this as proof that while our simple game-theoretic model captures the key financial incentives the players face, concerns about truth, fairness, or other aspects remain and do affect behavior, albeit to a small extent.

while it could be as well presented as a deposit since the fee is payable only in case of an unsuccessful ADR filing. The framing as a deposit could further strengthen the effect, which can be tested in the future. Naturally, the size of the fee influences the kind of cases that are then referred to as the ADR. Thus changing the fee might potentially lead to lower or higher success rates due to the self-selection of cases by content creators. Similarly, the rate might be higher if the blocked benefits from the disputed content at hand are further increased. Thus high-value content could follow a different trajectory than low-value content.

It is important to note that the low rates of complaints in real world could have at least four explanations: (1) the content that is blocked is not worth the cost of complaints, (2) the content creators are intimidated by the takedown because they think that providers and notifiers have superior knowledge of the law, (3) the content creators give up as they anticipate that providers will not reconsider due to legal risk and thus do not even bother complaining or (4) the content creators are actually worried about possible retaliation from the side of the providers (e.g. shadow banning by the provider's algorithm).

Our paper tackles the third situation. In particular, our research does not resolve the cases when the value of the blocked content is lower than the cost of complaining. It equally does not capture other than the financial motivation behind the decisions to dispute or not dispute the takedown. However, we believe that the other situations should be equally explored in future research. Interventions like the provision of better information about aggregate error rates, the use of personalized explanations, or community-driven peer-support might be tested as possible solutions. Future work could also explore potential strategic interactions between parties, for instance, the extent to which the content creators self-censor their creativity due to the risks of being blocked. That being said, in our view, the main driver of content creator's apathy today is the underlying lack of effective remedy, which we correct by introducing the ADR. The cases involving low value content, or intimidation can be also resolved by building collective redress tools on top of the ADR (e.g. an ability of users to cheaply refer cases to charities and NGOs).

We should also note that our ADR was designed to be self-sustainable. This is why the fee is not negligible. However, this feature can be changed if some third party, e.g., an NGO or a government, finances entirely or partly such complaints. In terms of policy options, there are at least two ways how to implement the proposed changes to the notice and takedown system. The first possibility would be to legislate an ADR as an option and incentivize providers to use it (e.g., by emphasizing the risk-free phase after the decisions are made). The second possibility would be to force such ADR mechanisms in a form of regulation. The third option is to substitute non-state ADR for public authorities or any other type of self- or co-regulatory bodies that would resolve the disputes. The legal nature of the dispute resolution body is less important than the financing design of such dispute resolution mechanisms. Namely, even if they are operated by the state, they have to preserve the fee payable by providers when they fail to defend the removal of content.

The proposal for the upcoming Digital Services Act in the European Union, which will update the law in the area, adopts the mandated version of our policy in its Article 18.[31] The provision establishes certification of non-state out-of-court bodies that can resolve content disputes and issue binding decisions. The content creators who win the disputes shall have their content reinstated and be reimbursed "for any fees and other reasonable expenses that the recipient has paid or is to pay in relation to the dispute settlement". Thus, the reimbursement is flexible enough to accommodate the self-sustaining fee structure of ADR bodies that will also act as an incentive for providers to avoid making mistakes. The provision thus illustrates how our model can be translated into a policy. Our paper provides evidence for its potential positive effects on the notice and take-down practice. If the provision is eventually adopted, there is room for further research to explore how to best calibrate the fees of such ADR for different types of content (e.g., hate speech, copyright infringement, etc.). Naturally, experimental work like ours cannot exactly calibrate the fees for all types of content. Each area of content and type of service has different dynamics. Calibration of fees could be best studied by field experiments. Perhaps most importantly, we show the usefulness of designing counter-incentives as a response to over-compliance of private actors who asked to enforce tasks for the state.

Finally, as any experimental project, ours too rests on a few modeling assumptions. First, the content creators and providers are rendered risk-free after any ADR decision. Second, in our set-up, the providers themselves are never hurt by over-compliance, only the content creator is. Note that in reality, for some types of content, providers could be hurt along with their content creators, and thus have stronger business counter-incentives to engage in a better quality of the review of notifications. We have already mentioned the example of search engines and the right to be forgotten as a potential candidate. Third, our set-up does not allow the provider to influence the quality of notifications that it receives; rather, both states of the world are equally likely. Policy-makers can also address this part of the enforcement chain. Fourth, providers in the baseline are not allowed to buy more time to review the notifications. Fifth, the benefit of content creators that are being blocked by the takedown is always higher than the costs of complaining to the provider. Sixth, the cost of complaining before the provider and ADR mechanism, when taken together, is higher than the blocked benefits. Seventh, the providers generally have sub-optimal time to evaluate the notifications received. And eighth, the interactions between content creators and providers are always separated, which means that they are not able to personalize their responses against each other (e.g., to block the content creator). However, as the recently proposed Article 18 Digital Services Act demonstrates, the majority of our assumptions are realistic and usually can be addressed and incorporated in a policy change *along* with our solution.

6. Conclusion

In an experiment, we show that an independent ADR mechanism can help mitigate social costs of over-blocking by companies once they are delegated tasks of removing illegal content from their services. Our design of such a mechanism significantly reduces over-compliance by providers. This occurs in three steps: first, the initial decisions by providers exhibit a substantially lower bias in favor of *takedown* because of counter-veiling incentives, second, the providers are more willing to reconsider their decisions when the content creators complain to them, and third and most importantly, the content creators view the ADR as an effective remedy and actually use it to resolve the remaining disputes. The existence of ADR leads to higher accuracy of provider's decisions, particularly in cases that are not too complicated. It also increases the content creator's profits. Overall, we significantly improve the accuracy of the delegated enforcement system, which benefits the business ecosystem and freedom of expression. We further show that our proposed policy serves as a tool to redistribute profits between the affected parties without decreasing the total profits (despite very conservative modeling assumptions).

Since we published the initial version of the paper, our solution was adopted by the European Commission as a part of its major overhaul of the digital services regulation in Europe. Our paper thus

[31] See Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC

provides empirical evidence for the potential positive effects of the proposed policy and identifies factors to consider when operationalizing it in practice. This, we hope, will contribute not only to the improvement of notice and take-down but also of other types of delegated enforcement policies.

Declaration of Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Instructions

A.1. Baseline

Welcome to this experiment. Note that you are not allowed to communicate with other people in this room or anywhere else by any means (e.g., talking, sending text messages, tweeting,.). As a thank you for showing up on time, you will receive 4 euros that will be added to your earnings at the end of the experiment.

This experiment consists of two parts. You will now receive instructions for part 1.

Part 1

You will now be matched with one other person in the room whose identity will remain anonymous to you. This person's decisions will, together with your own actions, determine your payoff. You will not change partners at any point during this part.

You will play for a total of **3 sets**, and each set consists of **5 decisions**. At the end of the experiment, one of these sets will be randomly chosen, and you will be paid your earnings from this set in Euro at an exchange rate 5 experimental dollars = 1 Euro.

At the beginning of every set the computer will determine which one of you will be player A and player B. Both players start every round in every set with an endowment of **10 experimental dollars**.

In every round, player A will be asked **whether a displayed maze has a solution** (Figs. A.1, A.2, A.3).

A maze might, or might not have a solution. Please have a look at Figs. 2 and 3 to see what is meant by having a solution.

Therefore, in each round, the correct answer is either YES or NO.

Player **A is never penalized for answering YES**, even if that answer is incorrect. However, if player A answers NO and this is incorrect (so, the maze actually has a solution), player A faces a 10% chance of losing all his/her earnings from that set of decisions.

So, if player A provides 1 incorrect NO answer, the risk is as stated above. For more mistakes within a set, they add up as follows: .

1. 2 mistakes: 20% chance to lose everything
2. 3 mistakes: 30% chance to lose everything
3. 4 mistakes: 40% chance to lose everything
4. 5 mistakes: 50% chance to lose everything

Remember, there are 5 decisions to make in every set, and only incorrect NO answers carry a risk. (Incorrect YES answers do not affect player A.).

Player B is also shown the puzzle and has twice as long to examine it. However, player B is not asked whether a solution exists.

Once player A has made their decision, player B is informed about it. **If player A answered NO, then regardless of whether this is true, player B is unaffected**. If player A answered YES, then regardless of whether this is true, player B's earnings are reduced by 4 experimental dollars.

So, YES answers (even if incorrect) carry no risk of losing earnings for player A, and NO answers (even if incorrect) never decrease the earnings of player B.

Player **B can punish player A by paying a fee of 2, which will reduce player A's earnings by 1**. Player B has the option to punish player A after every A's decision, regardless of what that decision was.

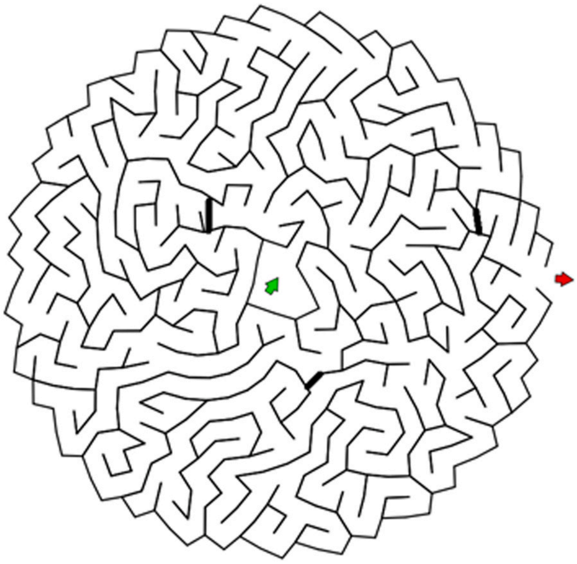

Fig. A.1. Example of a maze. Does it have a solution?.

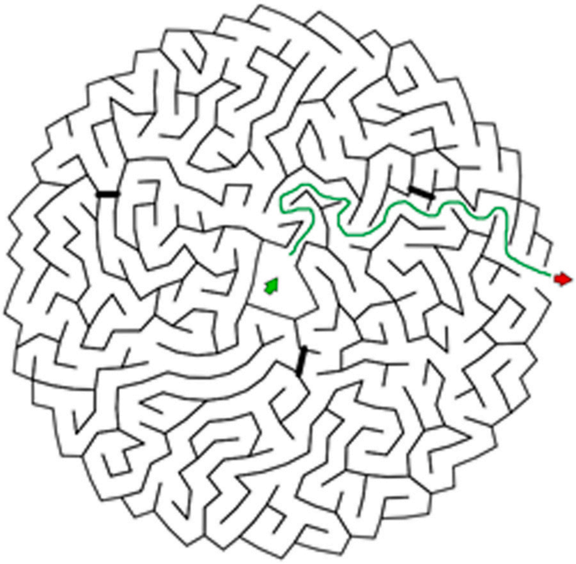

Fig. A.2. This maze has a solution.

However, there is only a 25% chance that the computer will implement the punishment: in 75% of the cases, the computer will ignore player B's request to punish. In those cases, player B still pays for punishment, but player A is unaffected.

Player A is informed in case punishment is implemented. If punishment is not implemented, player A does not know whether player B tried to punish him/her or not.

If player A is actually punished, he/she can re-evaluate the answer to the puzzle problem and switch. It is then this answer that counts as the final answer for determining the payoffs to both players, including the risk of losing earnings to player A. So, a final YES answer always hurts B for 4 dollars and does not affect A, a final NO does not affect B at all but – if incorrect – affects the risk that A loses their earnings. This ends the round.

At the end of each set, both players are shown their earnings for the set. For player A the computer also shows how many incorrect NOs they accumulated, and whether they lost any earnings as a result. There is no feedback on the correct solutions to the puzzles at any point in the experiment.

Please complete the comprehension quiz below. There are nine statements; decide for each whether it is true or false. When you are finished, please raise your hand. Once everybody has answered all questions correctly, we will start the experiment.

1. I will remain matched to my partner throughout part 1 (all sets). (T/F)
2. Player A has to answer YES or NO to the question "does the maze have a solution". (T/F)
3. Player A is never penalized for answering NO, even if the correct answer is YES. (T/F)
4. Player B has more time than player A to solve the puzzle. (T/F)
5. Player B can punish player A either for answering YES or for answering NO. (T/F)
6. The cost of punishment to player B equals 2. (T/F)

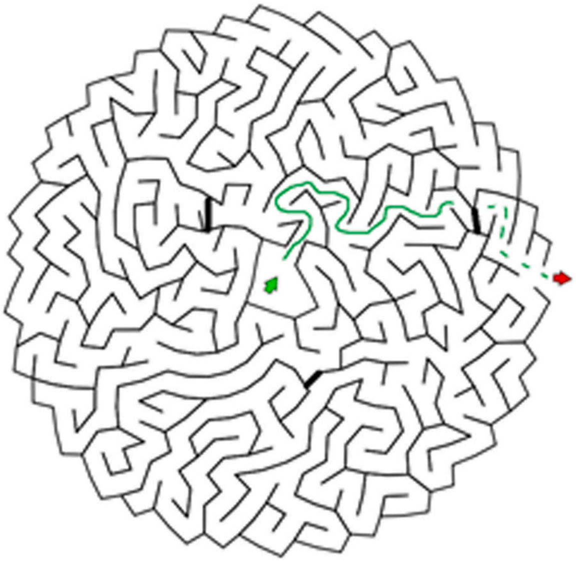

Fig. A.3. This maze does not have a solution.

7. Player A gets immediate feedback whether his/her YES/NO decision was correct. (T/F)
8. If Player B decides to punish player A, he/she knows for sure that he/she will have to pay the cost of punishment. (T/F)
9. If player B decides to punish player A, this punishment will take place with a probability of 75%. (T/F)

Part 2

In this part you will again play three sets of 5 decisions, with a **new partner** who will stay the same throughout this part. Again, one of these three sets will be randomly selected at the end for payment, and your earnings converted to Euro at a rate of 5 experimental dollars = 1 Euro (same rate as before).

In all other respects the game is the same as in part 1.

Again, please answer the short **comprehension quiz** below: .

1. I will have the same partner as in part 1. (T/F)
2. Player A knows when player B invested in punishment, even if the punishment is not implemented. (T/F)
3. If player A claims that the correct answer is NO, player B punishes, and punishment is not implemented, the respective earnings of these players will be: [show calculation]
4. If player A claims that the correct answer is YES, player B punishes, punishment is implemented, and player A does not change his decision, the respective earnings of these players will be: [show calculation]

A.2. ADR

Part 2

In this part you will again play three sets of 5 decisions, with a **new partner** who will stay the same throughout this part. Again, one of these three sets will be randomly selected at the end for payment, and your earnings converted to Euro at a rate of 5 experimental dollars = 1 Euro.

The game is the same as in part 1, with one difference:

If player B is unhappy with A's answer to the puzzle, he/she can challenge this outcome by submitting a complaint to a formal resolution body. This body can determine with 100% certainty whether YES or NO is the correct answer. Player B can submit this complaint in two different situations:

(a) Player B tried to punish player A but the punishment was not implemented
(b) Player B punished player A, the punishment was implemented, but player A did not change his/her answer

Submitting such a complaint costs player B **5 experimental dollars**. If he/she wins the dispute (the resolution body concludes that player A's answer was indeed incorrect), player A is forced to change his/her answer (so, switch from YES to NO or switch from NO to YES) and compensate the player B for having to complain by paying 8 experimental dollars to him/her. Since there is a certainty that this final answer of player A is correct, it carries no risk for player A.

Note that if the resolution body forces player A to switch from NO to YES, the negative effect of a YES answer still applies to player B: he/she gets reimbursed for complaining (8 dollars), but loses the 4 dollars he/she loses with every YES answer. If player A is forced to switch from YES to NO, player B gets not only the reimbursement of 8 dollars but also the damage of 4 is reversed (so, he/she gets 12 back in total).

If player B loses the dispute (so, player A's answer was correct), nothing further happens.

The resolution of the complaint ends the round.

Again, please answer the short **comprehension quiz** below:

1. I will have the same partner as in part 1. (T/F)
2. If player A claims that the correct answer is NO, player B punishes (punishment is not implemented), player B submits a complaint to the resolution body and wins the dispute, the respective earnings of these players will be: [show calculation]
3. If player A claims that the correct answer is YES, player B punishes (punishment is implemented), player A does not change his decision, player B submits a complaint to the resolution body and wins the dispute, the respective earnings of these players will be: [show calculation]
4. Player B cannot complain to the resolution body if he did not try to punish player A. (T/F)

Appendix B. Robustness checks

B.1. Baseline vs. ADR: first 15 rounds

In this subsection, we provide evidence that our treatments are comparable, i.e., that the subjects behaved identically in the first half of the experiment that was the same for both treatments. See Table 13 for details.

B.2. Within-group analysis

We include a within-subject analysis of how behavior of the subjects in the ADR treatment changes as they switch from baseline (first 15 rounds) to ADR (last 15 rounds). All our results remain unchanged with one exception: the introduction of ADR leads to fewer initial takedown mistakes, more punishment, more switching after punishment, lower profits of providers, higher profits of creators, and **insignificantly more correct final outcomes**; we believe this is because the number of mistakes made by providers increases with experience (as we saw in the analysis of the baseline in section 5.1), and so now the treatment needs to show an improvement compared to a relatively more favorable status quo Table 14.

Table 13
First 15 Rounds: Baseline vs. ADR.

	Baseline	ADR	p-value	Test statistic
Initial *takedown* decisions	59%	67%	0.0931	1.69
Punishment	12%	13%	0.7989	0.27
Switches after punishment	21%	10%	0.4509	-0.90
Profits providers	9.8	9.9	0.4534	− 0.77
Profits creators	7.4	7.1	0.1008	− 1.65

The table compares the behavior of players and the game outcomes in the (first 15 rounds of) the baseline and ADR treatment. In these 15 rounds, all subjects played under the baseline specification. All p-values are from the Wilcoxon rank-sum test conducted on the level of independent observations (N = 35; N = 26 for switches).

Table 14
Within-subject Analysis.

	Baseline	ADR	p-value	Test statistic
	(1–15 rounds)	(16–30 rounds)		
Initial *takedown* mistakes	72%	50%	0.0051	2.70
Punishment	13%	22%	0.0001	− 3.43
Switches after punishment	10%	58%	0.0078	− 2.75
Profits providers	9.9	8.2	0.0000	3.60
Profits creators	7.1	8.2	0.0004	− 3.24
Correct final outcomes	70%	75%	0.0941	− 1.69

The table compares the behavior of players and the game outcomes in the first 15 vs. last rounds of the ADR treatment averaged over independent observations. All p-values are from the Wilcoxon sign-rank test conducted on the level of independent observations (N = 17; N = 12 for switches).

B.3. Matching group size

In this subsection, we exclude the two matching groups that were due to a computer error of size 8 rather than 4, and that played only 10 periods in the ADR treatment. All our main results are robust to this change Table 15.

Table 15
Between-subject Analysis without Larger Matching Groups.

	Original result	Revised result
Decrease incorrect initial decisions	(n.s.)	(n.s.)
Decrease incorrect final outcomes	✓	✓
Decrease profits providers	✓	✓
Increase profits creators	✓	✓

The table compares our results from the (last 15 rounds of the) ADR treatment with and without the two matching groups made larger by accident. All evaluations of statistical significance are from the Wilcoxon rank-sum test conducted on the level of independent observations comparing the ADR treatment to baseline. A ✓denotes a significant increase or decrease, n.s. indicates a result not statistically significant.

B.4. Easy mazes classification

We noticed that not all mazes from the same level of difficulty are equally time-consuming; as a check, we ran an online survey with 48 respondents to see if students drawn from the same population as our subjects are able to solve all "easier" puzzles at higher rates than the "difficult" puzzles. This is important for knowing which information setting we are more likely to operate under. Additionally, we also provide evidence that (at least on average), content creators are more likely to know the correct solutions, i.e., know the "true state of the world".

The students were randomized into four groups: the puzzle ordering (baseline first or ADR first) was crossed with time availability (15 or 30 s time limit). Once students completed their first 15 puzzles, they were allowed to continue with another 15 in another treatment setting. The top performer from each of the four treatments was paid 10 Euro as a thank you for participation.

Unlike the experiment, the survey allowed respondents to choose 'I do not know' rather than pick either *takedown* or *keep*. We interpret such answers as possible signals that the puzzle is difficult since incorrect answers were punished, but 'I do not know' answers were not.

For this analysis, we use only the first set of 15 puzzles the students solved, since very few chose to also complete the second batch, and those results would then likely be plagued by self-selection.

In the strictest sense, if we want to talk about perfect information, the mazes should be so easy that every player can find the correct answer, in both time availability conditions. If we thus use this criterion, that 100% of all survey takers, in both provider and creator roles, we are left with 3 mazes (all from the first 15 rounds of the experiment) that can be classified as "easy". (Reassuringly, all three were from the easier level of difficulty.).

Unfortunately, since we do not have "easy" puzzles of this type in the second half of the experiment, we cannot check our main results using this definition. We, therefore, propose as an alternative definition of "easy": *The majority (> 50%) of subjects is able to solve the maze correctly, regardless of whether they played in the provider or creator role.* This seems reasonable if we accept that people can sometimes make mistakes by accidentally clicking, or might get distracted when filling in a survey and decide to skip a question. If we do so, the number of

Table 16
Easy vs. Difficult Puzzles Revisited.

	Baseline		ADR	
	Easy	Difficult	Easy	Difficult
Decisions correct	73%	54%	85%	58%
Incorrect *takedown* punished	27%	19%	90%	66%
Incorrect *takedown* ADR complaint	–	–	89%	96%

The table provides a comparison of the player's behavior in the (last 15 rounds of) baseline and ADR treatment. The table shows the percentage of providers whose initial decision was correct, the percentage of creators who punished an unjust *takedown*, and the percentage of creators who complained to the ADR when facing an incorrect *takedown*.

Table 17
Difficult Puzzles: Behavior of Providers Revisited.

	ADR	
	Easy	Difficult
Initial takedown decisions	47%	49%
▓▓...of which incorrect	23%	43%

The table compares the initial decisions taken by the providers in the (last 15 rounds of) the ADR treatment for easy and difficult mazes.

"easy" mazes increases to 10 (7 in the first half of the experiment, 3 in the second half). Of these, two (1 in each half of the experiment) were previously classified as difficult.

Using this alternative classification, we revisit our previous results.

As Table 16 and Table 17 show, the behavioral patterns are by and large the same as with our initial definition, with two minor differences: first, with this classification, we see a greater share of initially correct answers for easy puzzles, and second, fewer incorrect takedown decisions for easy puzzles (as we should, since we only define puzzles as easy if we have empirical evidence many people can solve them).

B.5. Do creators know more than providers?

Finally, we also look at whether having additional 15 s to solve our mazes is helpful to the subjects, as our initial results would seem to suggest. If so, it would imply that the creators are more likely to be informed than the providers.

In the aggregate, we find that our subjects score significantly better on the mazes when they had more time to solve them (Wilcoxon ranksum test p-value of 0.0312), but due to our small sample size, we cannot extend this conclusion to either treatment if analyzed separately.

We therefore cautiously conclude that while more time does help the players, the additional 15 s do not seem to provide an overwhelming informational advantage to the content creators. More thorough research on cases of information asymmetry is warranted.

Appendix C. Theory and proofs

In this section, we provide the derivation of the theoretical predictions for the three baseline types of games we discuss in the paper. Throughout, we assume that both players start with a prior belief that both states of the world ('takedown is correct' or 'keep is correct') are equally likely, players only care about their own monetary payoff, and players are risk-neutral. For easier reading, all actions players take are written in *italics* and we do not discuss actions taken by players in subgames that follow a strictly dominated strategy, as these do not affect the equilibrium outcome.

C.1. Prelude: dominant strategy of the content creator

First, as can be easily verified in the game tree, notice that regardless of the treatment or information setting, the content creator's dominant strategy is to *not punish* a *keep* decision by the provider. This is because such a decision yields the maximum possible payoff to the creator, with no risk, no matter the true state of the world.

Second, in the baseline treatment, the content creator will *not punish* a *takedown* decision, regardless of the information setting. By *not punishing* he can guarantee a payoff of 6, while by *punishing* he can earn up to 8 if punishment gets implemented and the provider responds in the most favorable way for the creator. Notice, however, that even if the provider always reconsidered his decision when punished, given how unlikely punishment implementation is, this strategy does not pay off for the content creator:

$$0.25*8 + 0.75*4 = 2 + 3 < 6$$

Third, in the ADR treatment, it is always optimal to *punish* an incorrect *takedown* decision, because either the provider will *switch* himself, or the *ADR* can be *invoked* and the content creator compensated. In the 'takedown correct' state of the world where the provider chose the *takedown* decision, the content creator will *not complain to the ADR* (as this yields negative payoff), and even if *punishment* always resulted in the provider *switching to keep*, this – just like in the baseline case – cannot happen often enough to justify the punishment cost.

Importantly, for the content creator to apply this strategy, he needs to know the true state of the world. If that is not the case, and so, we assume that neither player knows the true state of the world, *punishing* and *complaining to the ADR* yields a strictly lower expected payoff than *not punishing*, even in the most optimistic scenario where the provider always *switches* to 'keep' when punished and the ADR - in accordance with the creator's prior - rules in favor of the creator 50% of the time. The expected payoff is then only 5.75.

$0.5^*(0.75^*(-1) + 0.25^*8) + 0.5^*(0.75^*11 + 0.25^*8) = 5.75 < 6$

In the other possible scenario, the provider *not switching* when *punished*, this also yields a strictly lower payoff for the content creator:

$0.5^*(0.75^*(-1) + 0.25^*(-1)) + 0.5^*(0.75^*11 + 0.25^*11) = 5 < 6$

Note that the provider cannot condition his switching behavior on the state of the world, since by assumption it is unknown. It is likewise straightforward to verify that only *punishing* and *not following up with an ADR request* is not a profitable strategy either.

Taken together, the dominant strategy of the content creator can be easily summarized as follows:

- Do *not punish* any *keep* decision
- In the baseline, do *not punish* any *takedown* decision
- In the ADR, if you do not know the true state of the world, do *not punish* any *takedown* decision
- In the ADR, if you know the true state of the world, *punish* (and, if needed, *complain to the ADR*) in response to an incorrect *takedown* decision

Knowing what the content creator's dominant strategy is, we examine the six key cases for our game:

C.2. Game with perfect information

Here we assume that the puzzle in question is so easy that both players are able to determine the correct solution (i.e., the true state of the world) with certainty.

C.2.1. Baseline

In our baseline specification, the sequential equilibrium outcome depends on the (here: perfectly observable) true state of the world. When the state of the world is 'takedown is correct', the unique equilibrium outcome is for the provider to always *takedown*, and the content creator to *not punish*. In the 'keep is correct' state of the world the provider can pursue any linear combination of his two actions, (*takedown, keep*), while the content creator always responds with *not punish*.

Looking at the provider, in the 'keep is correct' state any action yields the same equilibrium payoff: no decision is *punished*, and since a *keep* decision is correct, the payoff of 10 is safe (i.e., does not yield any risk for the provider).

In the 'takedown is correct' state, *takedown (and no reconsideration)* is the dominant strategy for the provider, as it yields the maximum safe payoff. (We have already shown that the provider will not face *punishment*.).

C.2.2. ADR

The introduction of an ADR essentially simplifies the above analysis. It enables the content creator to enforce the correct response of the provider to a 'keep correct' state of the world, while the provider's self-interest ensures correct responses to a 'takedown correct' state. The unique equilibrium outcome is thus the provider matching his response with the state of the world (i.e., *takedown* if 'takedown correct' and *keep* if 'keep correct'), and the content creator *never punishing (or complaining to the ADR)*.

Consider the provider. Given the content creator's dominant strategy, it does not pay off to err on the *takedown* side, as all such mistakes will be punished (and, if needed, *taken to the ADR*). Such costs are avoidable, as the provider can safely stick with *keep* whenever 'keep is correct' is the true state of the world. Likewise, it is not profitable to err on the *keep* side, since those decisions carry a risk of losing earnings. Therefore, the best the provider can do is to always submit the correct decision and avoid all *punishment* altogether.

C.3. Game with informed content creator

Here we assume that the puzzle is of medium difficulty: the content creator has enough time to solve it, but the provider does not. Hence, the provider relies on his prior, while the content creator knows the true state of the world with certainty.

C.3.1. Baseline

In contrast to the game with perfect information, the state uncertainty on the part of the provider results in a unique equilibrium outcome: always *takedown*, to which the response is *not to punish*, regardless of the true state of the world.

First, notice that nothing changes from the content creator's perspective compared to the full information case.

However, the provider now does not know the true state of the world and is thus choosing between a safe *takedown* option that yields 10, or a risky *keep* option that yields 10, and in 50% of the cases is expected to be incorrect and carry a risk of losing earnings. Clearly, the dominant strategy is to always choose *takedown*.

C.3.2. ADR

Since the provider can only form an expectation whether his answers are going to be correct, he needs to find a balance between higher payoffs associated with the *keep* strategy, and safe, lower payoffs associated with the takedown strategy. This, combined with the content creator's strategy that is unchanged from section C.2.2, results in an equilibrium outcome that over the course of the five periods, the provider chooses *takedown* exactly once and *keep* exactly four times. *Takedown* is challenged (*punished* or *an ADR complaint is made*) only if it is incorrect; *keep* decisions are *never punished*.

The proof for the provider follows two steps: First, the provider forms an expectation of how much he can earn by choosing *takedown*. By assumption, *takedown* is the correct action 50% of the time (and this is the provider's prior as well). In these cases, the provider's decision will not be challenged, and he earns a safe return of 10. When *takedown* does not match the true state of the world, the content creator *punishes* the provider. In 25% of such cases, punishment is implemented. Since the provider knows that it is only profitable to *punish* when *takedown* is incorrect, he can safely switch his response to *keep*, resulting in a safe payoff of 10. However, in 75% of the cases, nature does not implement punishment, the provider cannot revise his answer, and the content creator ends up *complaining to the ADR*. This results in a loss for the provider, as he ends up with a safe payoff of 2. Taken together, the expected payoff from a *takedown* action equals:

$$0.5^*10 + 0.5^*(0.25^*10 + 0.75^*2) = 7$$

Second, starting from the fifth round, we can formulate the provider's problem using E for the total accumulated earnings up to that point, and n for the risky *keep* answers (i.e., those not following a revision) taken in the rounds prior. The provider prefers to make a risky *keep* decision if:

$$(E + 10)^*\left(1 - \frac{n+1}{20}\right) \geq (E + 7)^*\left(1 - \frac{n}{20}\right)$$

Which simplifies to:

$$50 \geq E + 3n$$

It is easy to verify that this condition holds for $n \leq 3$, but does not hold in case of $n = 4$: at that point, the provider has already gambled on *keep* four times, earning himself 40 and exposing himself to a great amount of risk. At that point, it is optimal for the provider to stick with *a takedown* as a safe option.

Using backwards induction, one can show that: .

- Assuming the provider opts for extextittakedown in the final period, it is always optimal to play *keep* in the previous rounds.
- Assuming the provider opts for *keep* in the final period, he will find it optimal to play *takedown* exactly once in the previous rounds, but its timing does not matter, as all strategies yield the same expected payoff
- The actually realized payoff when playing *takedown* does not affect follow-up decisions
- In every period, the condition for the provider to prefer playing *keep* rather than *takedown* can be simplified to $4 + 7T + 13K \geq E + 3n$, where T and K refer to the number of times *takedown* and *keep* strategies have already been played (starting from period one, as of the current period).

C.4. Game of no information

In this setup, we assume that the puzzle in question is so difficult that neither player gains any insight by studying it, and so neither player updates their prior of the probability of takedown being the correct answer.

C.4.1. Baseline

The unique equilibrium outcome is *takedown* followed by *not punish*, for both states of the world.

Notice that the provider's incentives and therefore also the optimal strategy are the same as in Section C.3.1, while *punishment* is still not profitable for the content creator, and so it never takes place.

C.4.2. ADR

Interestingly, this variant of the game has the same unique equilibrium outcome as its baseline counterpart: *takedown* followed by *not punish*, for both states of the world.

Recall that the content creator will *not punish* when he does not know the true state of the world; hence, the provider chooses a *takedown* as opposed to the *keep* option since it guarantees safe earnings of 10.

References

Bar-Ziv, S., Elkin-Koren, N., 2018. Behind the scenes of online copyright enforcement: empirical evidence on notice & takedown. Conn. Law Rev. 50 (2), 339–386.

Bridy, A., Keller, D., 2015. U.S. copyright office section 512 study: comments in-response to second notice of inquiry.

Christie, A., 2002. The ICANN domain-name dispute resolution system as a model for resolving other intellectual property disputes on the internet. J. World Prop. 5 (1), 105–118.

Dara, R., 2011. Intermediary liability in India: chilling effects on free expression on the internet. ⟨http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2038214⟩.

Erickson, K., Kretschmer, M., 2018. What motivates takedown of user-generated content by copyright owners? Evidence from the removal of music video parodies on youtube. J. Intellect. Prop. Inf. Technol. E-Commer. Law (JIPITEC) 9 (1), 75–89.

Farrand, B., 2013. Regulatory capitalism, decentered enforcement, and its legal consequences for digital expression: the use of copyright law to restrict freedomof speech online. J. Inf. Technol. Polit. 10 (4), 404–422.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Exp. Econ. 10 (2), 171–178.

Husovec, M., 2016. Accountable, not liable: injunctions against intermediaries. TILEC Discussion Paper No. 2016-012.

Husovec, M., 2021. (Ir)Responsible legislature? Speech risks under the EU's rules on delegated digital enforcement. Working paper.

Kaye, D.A., 2019. Speech police: the global struggle to govern the internet. Columbia Global Reports.

Kim, B., Kim, J.Y., 2017. The economics of the right to be forgotten. J. Law Econ. 60 (2), 335–360.

Klonick, K., 2017. The new governors: the people, rules, and processes governing online speech. Harv. Law Rev. 131 (6), 1598–1670.

Klonick, K., 2020. The facebook oversight board: creating an independent institution to adjudicate online free expression. Yale Law J. 129, 2418–2499.

Kur, A., 2002. Udrp - a study by the Max Planck Institute for foreign and international patent, copyright and competition law. ⟨https://www.zar.kit.edu/DATA/projekte/udrp_705937a.pdf⟩.

Nas, S., 2004. The multatuli project isp notice and take down.

Perel, M., Elkin-Koren, N., 2017. Black box tinkering: beyond transparency in algorithmic enforcement. Fla. Law Rev. 69, 181–221.

Randall, M.H., 2016. Freedom of expression in the internet. Swiss Rev. Int. Eur. Law 26 (2), 235–254.

Seng, D.K.B., 2014. The state of the discordant union: an empirical analysis of dmca takedown notices. Va. J. Law Technol. 18, 375–473.

Seng, D.K.B., 2015. Who watches the watchmen? An empirical analysis of errors in dmca takedown notices. ⟨https://ssrn.com/abstract=2563202⟩.

Tambini, D., Leonardi, D., Marsden, C., 2007. Codifying Cyberspace: Communications Self-regulation in the Age of Internet Convergence. Routledge, London.

Urban, J.M., Karaganis, J., Schofield, B., 2017. Notice and takedown in everyday practice. UC Berkeley Public Law Research Paper No. 2755628.

Urban, J.M., Quilter, L., 2006. Efficient process or 'chilling effects'? Take down notices under section 512 of the digital millennium copyright act. St. Clara Comput. High Technol. Law J. 22, 621–693.

Witt, A., Suzor, N., Huggins, A., 2019. The rule of law on Instagram: an evaluation of the moderation of images depicting women as bodies. Univ. NSW Law J. 42 (2), 557–596.