AN ECONOMIC MODEL OF ONLINE INTERMEDIARY LIABILITY

James Grimmelmann[†] *& Pengfei Zhang*[††]

ABSTRACT

Scholars have debated the costs and benefits of internet intermediary liability for decades. Many of their arguments rest on informal economic arguments about the effects of imposing different liability rules on online platforms. Some scholars argue that broad immunity is necessary to prevent overmoderation; others argue that liability is necessary to prevent undermoderation. These are economic questions, but they rarely receive economic answers.

In this Article, we seek to illuminate these debates by giving a formal economic model of online intermediary liability. The key features of our model are *externalities*, *imperfect information*, and *investigation costs*. A platform hosts user-submitted content, but it does not know which of that content is harmful to society and which is beneficial. Instead, the platform observes only the probability that each item is harmful. Based on that knowledge, it can choose to take the content down, leave the content up, or incur a cost to determine with certainty whether it is harmful. The platform's choice reflects the tradeoffs inherent in content moderation: between false positives and false negatives, and between scalable but more error-prone processes and more intensive but costly human review.

We analyze various plausible legal regimes, including strict liability, negligence, blanket immunity, conditional immunity, liability on notice, subsidies, and must carry, and we use the results of this analysis to describe current and proposed laws in the United States and European Union.

TABLE OF CONTENTS

DOI: https://doi.org/10.15779/Z38WP9T772

 The authors presented previous versions of this Article at the BTLJ-BCLT From the DMCA to the DSA symposium, the Freedom of Expression Scholars Conference, and the Cornell Tech Research Lab for Applied Law and Technology. They would like to thank the participants, Aislinn Black, Banks Miller, Elettra Bietti, Daphne Keller, Sanketh Menda, Clint Peinhardt, and Vitaly Shmatikov.

† Cornell Law School and Cornell Tech, Cornell University.

†† School of Economic, Political, and Policy Sciences, The University of Texas at Dallas.

I. INTRODUCTION

In the scholarly literature on intermediary liability, economic claims are common: for instance, that platform liability creates chilling effects;[1] that platforms do (or do not) have the right incentives to self-police;[2] and that platform liability creates a trade-off between protecting free speech and

1. *See, e.g.*, Felix Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293, 304 (2013).

2. *See, e.g.*, Danielle Keats Citron & Mary Anne Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, 2020 U. CHI. LEGAL F. 45, 52.

deterring abusive speech.[3] They are mostly informal policy arguments, not testable propositions. It is not clear when two claims conflict, or when they can coexist. Indeed, it is often not even clear whether two authors are making the same claim or different claims.

We do not propose to resolve any of these disputes. Instead, we aim to clarify the terms of the debate. In this Article, we recast arguments about online intermediary liability into a common language—the language of microeconomics. We give an economic model of online intermediary liability, with equations and diagrams. We see six significant benefits to legal scholarship from having such a model—benefits that in turn can help lead to better and more appropriately calibrated intermediary-liability law.

First and most fundamentally, modeling promotes communication. A suitable model can serve as a common framework for scholars to compare and contrast arguments. Our taxonomy of liability regimes reduces the at-times bewildering array of arguments about the proper scope of intermediary liability into a (we hope) orderly structure that makes it straightforward to see how different claims relate.

Second, modeling promotes intuition. A good model can bring out the consequences of a course of conduct or make plain why parties behave the way that they do. There are several common patterns in intermediary-liability law that have simple and vivid expressions in our model.

Third, modeling promotes visualization. We have attempted to provide a simple and memorable visual shorthand for every moving part in our model and every interesting effect of a legal regime. For example, we hope that even if you take nothing else away from this Article, you will have a clear visual sense for why a platform might either overmoderate or undermoderate even in the absence of liability.

Fourth, modeling promotes rigor. The process of writing down a model forces one to make one's assumptions explicit. Reasoning through a model's consequences requires a close examination of each claimed effect. In the course of working through our own model, we learned a lot about how arguments for and against intermediary immunity work, and this Article conveys some of what we have learned.

Fifth, modeling promotes proof. Given a set of explicit assumptions, it is possible to show rigorously whether particular conclusions follow. For

3. *See, e.g.*, Daphne Keller, *Toward a Clearer Conversation About Platform Liability*, KNIGHT FIRST AMEND. INST. (Apr. 6, 2018), https://knightcolumbia.org/content/toward-clearer-conversation-about-platform-liability.

example, we demonstrate that under our assumptions, strict liability consistently results in overmoderation. Of course, the real world is not required to comply with a proof about a model. But proofs like these make models more useful because they help pin down the predictions made by the model.

Sixth, modeling promotes empiricism. This is not an econometric Article; there are no datasets and very few numbers. But a model like ours helps identify the right econometric questions to ask. We hope that it provides a roadmap for future empirical work.

This Article has seven Parts, including this Introduction and a brief Conclusion. Part II surveys previous work in this space. Part III describes the model in formal detail. Parts IV and V analyze a variety of liability regimes in detail. Part VI shows how various current and proposed laws map on to those liability regimes. Part VII concludes and discusses possible extensions of the model.

II. BACKGROUND AND RELATED LITERATURE

At the most general level, this Article asks whether a regime of *online intermediary liability* or *online intermediary immunity* is economically optimal. A liability regime requires platforms to compensate the victims of harmful content posted by the platform's users. An immunity regime, by contrast, does not impose such liability on platforms—even when the users themselves might be held liable for posting the harmful content. There are numerous variations and hybrids of these two basic regimes, but this dichotomy is the fundamental legal and policy question of online intermediary law.

The literature on the economic analysis of law and the legal scholarship on intermediary liability are both immense. In the former, we draw particularly on the tradition of formalizing liability rules started by John Prather Brown,[4] and on the standard distinction between strict liability and negligence.[5] In this literature, the usual goal of the liability system is to promote efficiency, which means minimizing total social costs. The focus is on the effect of different liability rules on incentives for taking precautions to reduce risk. One key insight is that if only injurers influence risks, both strict liability and negligence can induce them to take optimal care.

4. John Prather Brown, *Toward an Economic Theory of Liability*, 2 J. LEGAL STUD. 323 (1973).

5. *See id.*; *see also* Steven Shavell, *Strict Liability Versus Negligence*, 9 J. LEGAL STUD. 1 (1980). *See generally* STEVEN SHAVELL, ECONOMIC ANALYSIS OF ACCIDENT LAW (2009) (summarizing literature).

The economic theory of liability underpins the product liability regime, where a manufacturer or seller might be held liable for harm caused by a defective or unsafe product.[6] This theory takes into account the relationship between liability, market price, and firms' profit-maximizing production. One of the key insights of this literature is that whether and how to impose liability depends on the characteristics of the product and the information available to consumers. For example, a strict liability rule would be more appropriate when it is difficult to test for product safety. In general, strict liability is more efficient than negligence, as it either results in prices that induce optimal production or induces consumers to purchase the optimal quantity. One of the key purposes of our model is to square this conclusion with the widespread argument in the legal literature that strict liability is particularly inappropriate for platforms.

The economic literature on online intermediary liability, however, is still in its infancy. The most detailed treatment is a student note by Matthew Schruers (now the president of the Computer and Communications Industry Association).[7] He analyzes four legal regimes—negligence, notice-based liability, strict liability, and immunity—in a model where the intermediary can vary its level of care. Although the moving parts in his model are different than ours, it is an illuminating analysis of the tradeoffs involved in imposing intermediary liability. His most trenchant insight is that notice reduces the effort required for the platform to achieve a given level of care,[8] an approach that informs our own information-based treatment of liability on notice. He also notes the essential parallel between liability on notice and strict liability and the chilling effect of strict liability on online speech.

In a more recent article, Xinyu Hua and Kathryn Spier extend the product-liability framework to a two-sided platform that enables interactions between sellers and buyers.[9] By either raising its prices or investing in screening, the

6. *See* Koichi Hamada, *Liability Rules and Income Distribution in Product Liability*, 66 AM. ECON. REV. 228 (1976); A. Mitchell Polinsky, *Strict Liability vs. Negligence in a Market Setting*, 70 AM. ECON. REV. (PAPERS & PROC.) 363 (1980); William M. Landes & Richard A. Posner, *A Positive Economic Analysis of Products Liability*, 14 J. LEGAL STUD. 535 (1985); A. Mitchell Polinsky & Steven Shavell, *The Uneasy Case for Product Liability*, 123 HARV. L. REV. 1437 (2010). *See generally* Andrew F. Daughety & Jennifer F. Reinganum, *Economic Analysis of Products Liability: Theory*, *in* RESEARCH HANDBOOK ON THE ECONOMICS OF TORTS 69 (Jennifer H. Arlen ed., 2013) (summarizing literature on economics of product liability).

7. Matthew Schruers, Note, *The History and Economics of ISP Liability for Third Party Content*, 88 VA. L. REV. 205 (2002).

8. *Id.* at 237–39.

9. Xinyu Hua & Kathryn E. Spier, *Holding Platforms Liable*, HKUST BUSINESS SCHOOL RESEARCH PAPER NO. 2021-048 (2022), https://ssrn.com/abstract=3985066 or http://dx.doi.org/10.2139/ssrn.3985066.

platform wants to keep safe sellers while deterring harmful sellers. They argue that whether to impose liability on the platform depends on whether the sellers are judgement-proof. If the sellers have deep pockets, then intermediary immunity is optimal. If the sellers are instead judgement-proof, then intermediary liability is necessary, and imposing residual liability on the platform improves social welfare.[10]

Intermediary immunity, as codified by § 230, has been credited for promoting the growth of the internet.[11] Many authors argue that § 230 was a response to concerns about the negative impact of lawsuits on online service providers, and that § 230 strikes a balance between free speech and safety.[12] Some scholars observe that platforms do moderate in the absence of any liability.[13] They argue that even intermediary immunity might lead to the over-removal of content by the platform (through user account termination, shadow banning, or "collateral censorship").[14]

At the same time, other scholars criticize the current shape of § 230.[15] Danielle Citron has argued that online platforms facilitate and amplify harassment and hate speech.[16] She and others argue that courts have given § 230 an overly broad interpretation and that this broad immunity provides excessively strong incentives to allow or encourage online materials to go unmoderated.[17] Consequently, the broad immunity fails to protect the victims

10. *See* Yassine Lefouili & Leonardo Madio, *The Economics of Platform Liability*, 53 EUR. J.L. & ECON. 319 (2022).

11. Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639, 650–57 (2013).

12. *See, e.g.*, Paul Ehrlich, *Communications Decency Act 230*, 17 BERKELEY TECH. L.J. 401, 411–13 (2002); Cecilia Ziniti, *Optimal Liability System for Online Service Providers: How Zeran v. America Online Got It Right and Web 2.0 Proves It*, 23 BERKELEY TECH. L.J. 583, 584 (2008); Matt C. Sanchez, *The Web Difference: A Legal and Normative Rationale Against Liability for Online Reproduction of Third-Party Defamatory Content*, 22 HARV. J.L. & TECH. 301, 317–19 (2008); Jeff Kosseff, *Defending Section 230: The Value of Intermediary Immunity*, 15 J. TECH. L. & POL'Y 123, 144–45 (2010).

13. *See, e.g.*, Eric Goldman, *Online User Account Termination and 47 U.S.C. § 230(c)(2)*, 2 U.C. IRVINE L. REV. 659, 670–71 (2012).

14. *See, e.g.*, Wu, *supra* note 1, at 296–97.

15. *See generally* Joel R. Reidenberg, Jamela Debelak, Jordan Kovnot & Tiffany Miao, *Section 230 of the Communications Decency Act: A Survey of the Legal Literature and Reform Proposals*, FORDHAM L. LEGAL STUDIES RSCH. PAPER NO. 2046230 (2012), https://ssrn.com/abstract=2046230 (surveying scholarly analyses of Section 230 and proposed reforms).

16. DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 61 (2014).

17. *See id.*; *see also* Jennifer Benedict, *Deafening Silence: The Quest for a Remedy in Internet Defamation*, 39 CUMB. L. REV. 475, 493 (2008); Colby Ferris, *Communication Indecency: Why the Communications Decency Act, and the Judicial Interpretation of It, Has Led to a Lawless Internet in the Area of Defamation*, 14 BARRY L. REV. 123, 136 (2010).

of online abuse with no recourse against the platforms, whose profit maximizing business models facilitate the harmful activities.[18]

There is substantial literature advocating the reform of § 230, but scholars disagree on the appropriate form of intermediary liability. Some papers seek to refine the scope of intermediary immunity by taking a multi-factor approach. For example, some authors suggest that courts should consider the level of editorial control exercised by the platform and the harm caused by defamatory statements when determining immunity in each case (though it is not clear how the court should weigh these different factors).[19] Other authors discuss strict liability for certain kinds of harms;[20] in particular, Nancy Kim suggests treating intermediary liability like product liability.[21] And some authors suggest applying criminal liability to sites that have been facilitating and profiting from illegal activities.[22]

Other scholars propose forms of conditional immunity. Danielle Citron and Benjamin Wittes have argued that the platforms should enjoy immunity only if they can prove that they took reasonable efforts to address online abuse, and lawmakers should specify the obligations for this duty of care.[23] Doug Lichtman and Eric Posner suggest a conditional immunity rule in which ISPs would be held liable for infringing content only if they fail to implement reasonable measures to prevent or deter infringement.[24] Caitlin Hall proposes

18. *See, e.g.*, Danielle K. Citron & Benjamin Wittes, *The Problem Isn't Just Backpage: Revising Section 230 Immunity*, 2 GEO. L. TECH. REV. 453 (2018); Ann Bartow, *Internet Defamation as Profit Center: The Monetization of Online Harassment*, 32 HARV. J.L. & GENDER 383 (2009).

19. *See, e.g.*, Jae Hong Lee, *Batzel v. Smith & Barrett v. Rosenthal Defamation Liability for Third-Party Content on the Internet*, 19 BERKELEY TECH. L.J. 469, 493 (2004); Vanessa S. Browne-Barbour, *Losing Their License to Libel: Revisiting § 230 Immunity*, 30 BERKELEY TECH. L.J. 1505, 1553–56 (2015).

20. *See, e.g.*, Mark MacCarthy, *What Payment Intermediaries Are Doing About Online Liability and Why It Matters*, 25 BERKELEY TECH. L.J. 1037, 1043 (2010) (discussing the payments industry as a source of moderation that responds to legal incentives).

21. Nancy S. Kim, *Imposing Tort Liability on Websites for Cyberharassment*, 118 YALE L.J. POCKET PART 115, 117 (2008).

22. *See* Shahrzad T. Radbod, *Craigslist—A Case for Criminal Liability for Online Service Providers*, 25 BERKELEY TECH. L.J. 597 (2010).

23. *See* Danielle Keats Citron, *How to Fix Section 230*, 103 B.U. L. REV. (forthcoming 2023); Citron & Wittes, *supra* note 18.

24. *See* Doug Lichtman & Eric A. Posner, *Holding Internet Service Providers Accountable*, 14 SUP. CT. ECON. REV. 221, 251–54 (2006). Lichtman and Posner focus on ISPs' role in "the creation and propagation of worms, viruses, and other forms of malicious computer code." *Id.* at 221. This is still a form of content moderation; the content at issue is harmful to computer systems.

an immunity conditioned on the display of rating labels that alert internet users of the credibility of information posted on the sites.[25]

Other authors compare the different immunity regimes of § 230 and § 512.[26] Those who highlight what they see as the advantages of the DMCA-style regime note the ability of the victims to prevent further harm through a notice-and-takedown system, the coordination of internet service providers (ISPs) in removing harmful materials, and the administrative ease of transition. On the other hand, those who worry about applying a notice-and-takedown system to all user-generated content highlight the possible chilling effects on free speech. One important line of work discusses graduated-response, in which ISPs issue warnings and impose penalties on users who engage in infringing behaviors and use the termination of service as a threat to induce lawful behavior.[27]

Other authors compare the intermediary liability regime in the European Union with that in the United States. Daphne Keller examines the relationship and the tension between online platforms' liability in Europe (e.g., the E-Commerce Directive) and the European Union's General Data Protection Regulation (GDPR).[28] Miriam Buiten, Alexandre De Streel, and Martin Peitz examine the European Union's Digital Single Market strategy, and they argue that the current E.U. liability framework is inadequate for dealing with the challenges of online content moderation.[29] They claim that the absence of "Good Samaritan" protection in the E.U. e-Commerce Directive creates

25. Caitlin Hall, Note, *A Regulatory Proposal for Digital Defamation: Conditioning § 230 Safe Harbor on the Provision of a Site "Rating,"* STAN. TECH. L. REV. N1 (2008).

26. *See* Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. ON TELECOMM. & HIGH TECH. L. 101, 102–04 (2007); Sarah Duran, *Hear No Evil, See No Evil, Spread No Evil: Creating a Unified Legislative Approach to Internet Service Provider Immunity*, 12 U. BALT. INTELL. PROP. L.J. 115, 118–33 (2004); Olivera Medenica & Kaiser Wahab, *Does Liability Enhance Credibility: Lessons from the DMCA Applied to Online Defamation*, 25 CARDOZO ARTS & ENT. L.J. 237, 256–62 (2007); Cyrus Sarosh Jan Manekshaw, *Liability of ISPs: Immunity from Liability Under the Digital Millennium Copyright Act and the Communications Decency Act*, 10 COMPUT. L. REV. & TECH. J. 101, 110–32 (2005); Jonathan Band & Matthew Schruers, *Safe Harbors Against the Liability Hurricane: The Communications Decency Act and the Digital Millennium Copyright Act*, 20 CARDOZO ARTS & ENT. L.J. 295, 296–319 (2002).

27. *See generally* Peter K. Yu, *The Graduated Response*, 62 FLA. L. REV. 1373 (2010); Annemarie Bridy, *Graduated Response American Style: "Six Strikes" Measured Against Five Norms*, 23 FORDHAM INTELL. PROP., MEDIA & ENT. L.J. 1 (2012); Rebecca Giblin, *Evaluating Graduated Response*, 37 COLUM. J.L. & ARTS 147 (2014).

28. Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU General Data Protection Regulation*, 33 BERKELEY TECH. L.J. 287, 351–61 (2018).

29. Miriam C. Buiten, Alexandre De Streel & Martin Peitz, *Rethinking Liability Rules for Online Hosting Platforms*, 28 INT'L J.L. & INFO. TECH. 139 (2020).

perverse incentives for platforms not to monitor online activity, thus undermining self-regulation.[30]

There is also a strong empirical literature on content moderation.[31] Collectively, this research suggests that platforms tend to have a bias towards over-removal.[32] In 2005, Jennifer Urban and Laura Quilter presented the first set of descriptive statistics on the notice-and-takedown process under DMCA § 512.[33] They found that corporations and business entities were the primary senders of notices, a majority of the notices were sent for competition purposes, one third of the notices were questionable regarding the validity of the copyright infringement claim, and very few individual users responded with a counter-notice.[34] In a follow-up study, Urban, Joe Karaganis, and Brianna Schofield emphasize the role of automation in sending complaints, and compare how the automated notices differ from the manual notices by small right holders.[35] More recently, Daniel Seng compiled a larger dataset and gave more detailed statistics, questioning the validity of many takedown notices, especially those generated by automated systems.[36] Meanwhile, Jonathon Penney surveyed 1,296 panelists with hypothetical scenarios on receiving a takedown notice, and his findings indicate some chilling effects of the policy.[37] Respondents broadly reported being less likely in future not only to share the same content again, but also to share content they themselves had created

30. *Id.* at 163–66.

31. *See generally* Daphne Keller & Paddy Leerssen, *Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM 220 (Nathaniel Persily & Joshua A. Tucker eds., 2020) (surveying information released by platforms and independent research).

32. Daphne Keller, *Empirical Evidence of Over-Removal by Internet Companies Under Intermediary Liability Laws: An Updated List*, STAN. L. SCH.: CTR. FOR INTERNET & SOC'Y (Feb. 8, 2021), https://cyberlaw.stanford.edu/blog/2021/02/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

33. Jennifer M. Urban & Laura Quilter, *Efficient Process or "Chilling Effects"? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA COMPUT. & HIGH TECH. L.J. 621 (2005).

34. *Id.* at 649–80.

35. *See generally* JENNIFER M. URBAN, JOE KARAGANIS & BRIANNA SCHOFIELD, NOTICE AND TAKEDOWN IN EVERYDAY PRACTICE (2nd version 2017).

36. Daniel Seng, *Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices*, 37 SANTA CLARA HIGH TECH. L.J. 119 (2020).

37. Jonathon W. Penney, *Privacy and Legal Automation: The DMCA as a Case Study*, 22 STAN. TECH. L. REV. 412 (2019).

(seventy-two percent).[38] Only thirty-four percent said they would send a counter-notice or challenge a takedown they believed was wrong or mistaken.[39]

III. AN ECONOMIC MODEL OF PLATFORM CONTENT MODERATION

There are two distinctive features of platform liability for harmful third-party content. The platform has *imperfect information* about which content is harmful and which is not, and content can have *positive externalities* not captured by the platform itself. These two features, taken together, mean that holding the platform liable for the harmful content it carries can go wrong. Because the platform cannot perfectly distinguish harmful from harmless content, and because it does not internalize the full benefits from the harmless content, the threat of liability can cause the platform to overmoderate, removing too much harmless content along with the harmful content.

A. THE MODEL

The first essential feature that makes intermediary liability distinctive is that a platform has imperfect information about the content that it hosts. Some content is harmful, and other content is not, but they look the same on first glance. A court decides whether a statement is legally defamatory after fact discovery, motion practice, and a trial; a platform does not have the time, the resources, or the power to conduct a full civil lawsuit on every post. A court awards damages in the fullness of time, on relatively complete information; a platform must act now, with radically incomplete information.

Formally, users submit discrete items x_1, x_2, x_3 ... of **content** to a platform. Each of these items is either **harmful** or **harmless**, and the platform can either **host** or **remove** each item. The essential feature of the model is that platform *does not know* whether each item is harmful or not. Instead, it observes the probability $\lambda(x)$ that item x is harmful, so it must make its hosting or removal decision under conditions of uncertainty.

38. *Id.* at 446–47.
39. *Id.* at 451.

Figure 1: Probability of harm

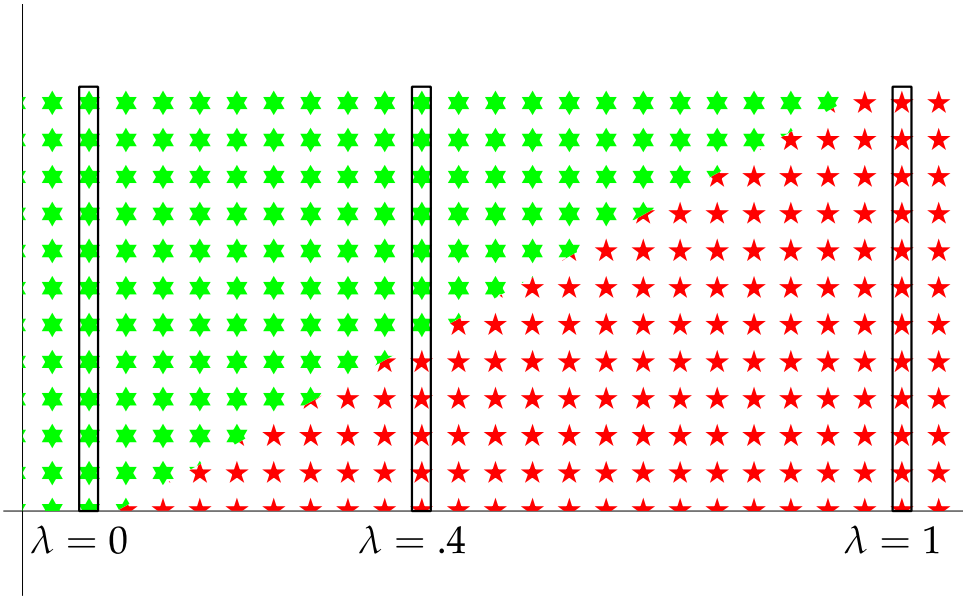


Figure 1 illustrates the platform's imperfect information. Think of the content presented to the platform as being divided into buckets. The platform knows what fraction (the probability $\lambda(x)$) of the content in each bucket is harmful (red five-pointed stars) or harmless (green six-pointed stars). But it does not know which specific items of content (individual stars) are harmful or harmless. If the platform hosts an item x, it has the following consequences:

- The **platform** receives some revenue $p(x)$.
- **Society** realizes some benefits $s(x)$.
- If the item is harmful, a third-party **victim** suffers harm $h(x)$.

If the platform removes the item, then the revenue, social benefits, and third-party harms are all 0.

Note that the social benefits of content $s(x)$ are known with certainty, and so are the harms $h(x)$ *if they happen*. Overall social welfare is therefore $s(x) - h(x)$ for harmful items, and $s(x)$ for harmless ones. Thus, the *expected* social welfare from hosting an item of content is $s(x) - \lambda(x)h(x)$: the known benefits minus the expected harms.

In general, $p(\cdot)$, $s(\cdot)$, $h(\cdot)$, and $\lambda(\cdot)$ could be arbitrarily complicated functions that account for an arbitrarily large number of features of each item of content. So while this expression is almost tautologically simple, it does not say much about how to draw useful lines between different kinds of content.

Therefore, we simplify the model by collapsing all content to a *single axis*. Imagine the content submitted by users to a platform arranged on a spectrum from worthwhile to worthless. At one end, the content is entertaining and informative—cat pictures and civics lessons. At the other end, the content is stomach-churning or worse—gross-out pictures and badly-written spam. A platform sets its moderation policy by deciding where along this spectrum to draw the line.

More formally, we assume that each item content falls within the one-dimensional interval from 0 to $x_{\max}$, where 0 is the "good" end and $x_{\max}$ is the "bad" end. Then as x increases:

- Content is less profitable to the platform: $p(x)$ decreases.
- Content is less beneficial to society: $s(x)$ decreases.
- The harm (if it happens) is fixed: h is a constant.
- Content is more likely to be harmful: $\lambda(x)$ increases.

We assume that $s(x) > p(x)$, i.e., all content has some positive spillover benefits for society that the platform does not capture.[40] We do not assume that $p(x) > 0$ or $s(x) > 0$: it is possible that some content is negative-value even if it is not harmful to third parties. (An example is spam, which is costly for the platform to host and has infinitesimal spillover benefits for anyone else.)

To make the model interesting, and to eliminate some annoying corner cases, we assume that the most innocuous content is profitable for the platform ($p(0) > 0$), beneficial to society ($s(0) > 0$), and known with certainty to be harmless, i.e., $\lambda(0) = 0$. Similarly, we assume that most problematic content is known with certainty to be harmful ($\lambda(x_{\max}) = 1$) and that harmful content is unambiguously bad for society, i.e., $h > s(x)$ for all x. These conditions ensure that some content is definitely good for society and some content is definitely bad for society, so that there is a real interest in treating them differently.

40. Any negative spillovers are separately accounted for by the harm h.

Figure 2: A one-dimensional model of moderation

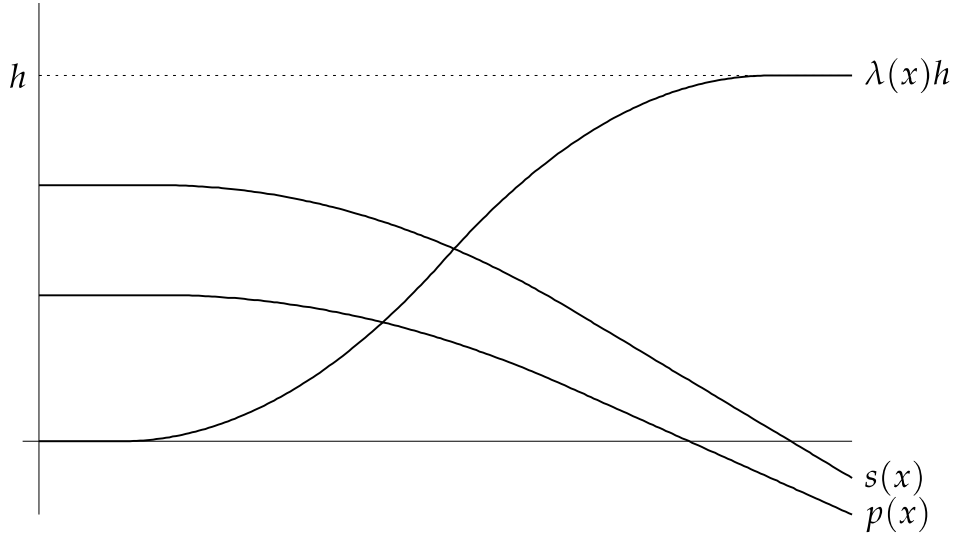


Figure 2 illustrates the essential model. The platform-revenue and social-benefit curves $p(x)$ and $s(x)$ start off positive and drop off. The expected-harm curve $\lambda(x)h$—the probability that content is harmful times the harm if it is—starts at 0 and rises to h.Given these assumptions, it is easy to see that content further to the left is always better ex ante and content further to the right is always worse. If $x_1 < x_2$, then x_1 is more profitable to the platform, better for society, and less likely to be harmful ex ante. Of course, if x_2 turns out to be harmless and x_1 is not, then x_2 might be better ex post, but from behind the veil of probabilistic ignorance, x_1 is the better prospect ex ante.

It follows that a rational moderator who is concerned with maximizing benefits and profits and minimizing harms will set a **moderation threshold** x^*. It will leave up all content x with $x < x^*$, and remove all content x with $x > x^*$. There is no circumstance under which it makes sense for the moderator to take down content x and leave up content y where $x < y$, because it would always be better to leave up x and take down y instead.

Any choice of x^* trades off false positives and false negatives. A low threshold means that more harmless content will be removed; a high threshold means that more harmful content will stay online. We tolerate some harmful content because it is indistinguishable ex ante from harmless content. The choice of x^* incorporates the moderator's judgments about the *acceptable risk of harm*.

This imperfect information is central to our model, and we believe that it is a pervasive fact of content moderation. While the users who upload content and the victims who are harmed by it may be in a better position to know whether content is harmful, platforms and regulators operate from a position of comparative ignorance.

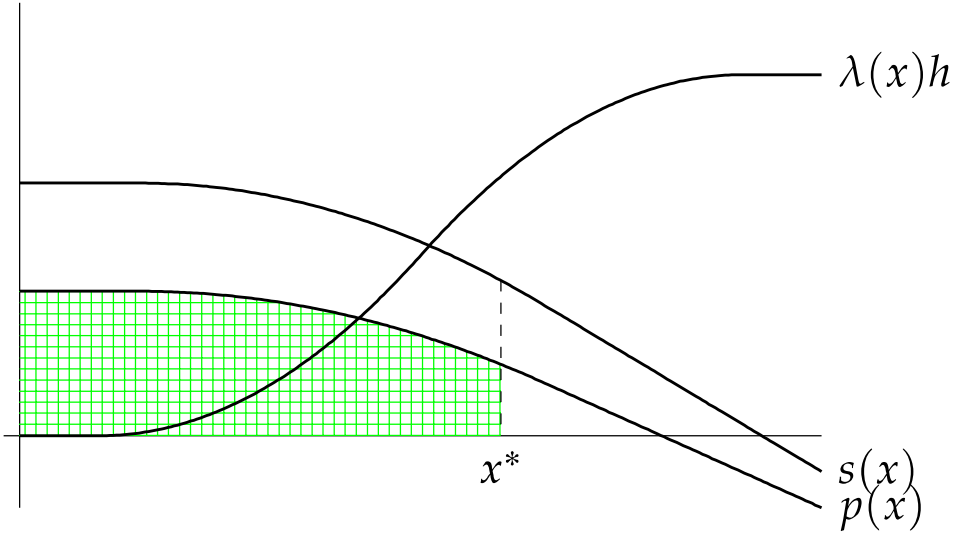
The overall harm here is a *statistical* consequence of a given choice of x^*. If the platform could perfectly distinguish harmful and harmful content, it could choose to host only the harmless content. (Indeed, we will shortly consider an extension of the model under which this distinction is possible, albeit at a cost.) But the point of the current model is that the platform cannot distinguish the two. A choice of x^* is a choice about the acceptable ratio of babies to bathwater.

Figure 2 is an abstract microeconomic diagram. Its purpose is to build qualitative intuition, not to be a scale model of anything specific. The x-axis is measured in abstract "units" of content. Think of each short interval along the axis as being occupied by an indefinitely large number of individual items of content. There are so many items, in fact, that we will treat the interval $[0, x_{\max}]$ as being effectively continuous; while content comes in distinct items, they are individually too small to be visible to the naked eye. Similarly, the y-axis is measured in abstract units of value. They could be dollars, or euros, or utils. Thus the values of the functions $p(x)$ and $s(x)$ and the constant h have the units of "value per unit of content," where, to repeat, a "unit" is made up of many individual items.[41]

We emphasize this point because it is important to remember that this diagram portrays the *marginal* platform revenue, social benefit, and third-party harm per unit of content. The value of $p(x)$ at a point x is the amount of additional revenue the platform will earn by hosting one additional unit of content at x—i.e., from increasing x^* by one unit. The value $p(x)$ is emphatically not the platform's *total* revenue from setting its moderation threshold to x.

41. The value of the function $\lambda(x)$ is a unitless probability, a number between 0 and 1.

Figure 3: Curves represent marginal value; areas represent total value



Rather, total profits, benefits, and harms are illustrated in Figure 1 (and in the numerous diagrams that will follow) by *areas*. For example, Figure 3 illustrates the platform's profits from setting its moderation level at x^*. At any given point, the vertical distance from the x-axis to the revenue curve $p(x)$ is the platform's marginal revenue from hosting the content at x. The platform's total profits are the area of the green checked region.[42]

B. SOCIAL WELFARE, PLATFORM PROFITS, AND BLANKET IMMUNITY

Now we are ready to use the model to draw conclusions about what the platform will do, and what the regulator wants it to do, which are not necessarily the same. We begin by asking what the socially optimal moderation level would be, and then consider whether the platform will set its moderation at that level. (Spoiler alert: no.)

The marginal social welfare from hosting content is the (known) social benefit from that content minus the (expected) harms, i.e.,

$$s(x) - \lambda(x)h. \tag{1}$$

42. In calculus terms, the platform's total profits are the *integral* of its marginal revenues, i.e.,

$$\int_0^{x^*} p(x)\,dx.$$

Another way to look at this expression is that if the platform hosts content at x, a fraction $\lambda(x)$ of that content will be harmful with value $s(x) - h$ per unit: benefits minus harms. Meanwhile, a fraction $1 - \lambda(x)$ will be harmless with value $s(x)$ per unit: all benefits and no harms. In other words, all of the content, harmful and harmless alike, generates benefits of $s(x)$, but only the harmful fraction $\lambda(x)$ also generates harms h.

Geometrically, the marginal social welfare from hosting content is the vertical distance between the benefit curve $s(x)$ and the expected harm curve $\lambda(x)h$. That value is 0 where the two curves cross.[43]

Call this point x_s, i.e., the **socially efficient moderation level**. It is defined by the equation

$$s(x_s) = \lambda(x_s)h.$$

For $x < x_s$, it is net beneficial to society for the platform to host this content. For $x > x_s$, it is net harmful to society. x_s is the point at which content crosses over from being net beneficial to net harmful. The regulator would prefer the platform to set its moderation level to x_s—i.e., to host content just up to x_s and then stop and take down everything else.

Observe how the value of x_s depends crucially on h. Rearranging the defining equation yields $\lambda(x_s) = \frac{s(x_s)}{h}$. The greater the harm h, the lower the probability $\lambda(x)$ of harm worth tolerating, and thus the lower the appropriate threshold of moderation.

43. By the intermediate value theorem, there is some value of x at which $s(x) - \lambda(x)h = 0$, so the curves do cross.

Figure 4: Optimal moderation

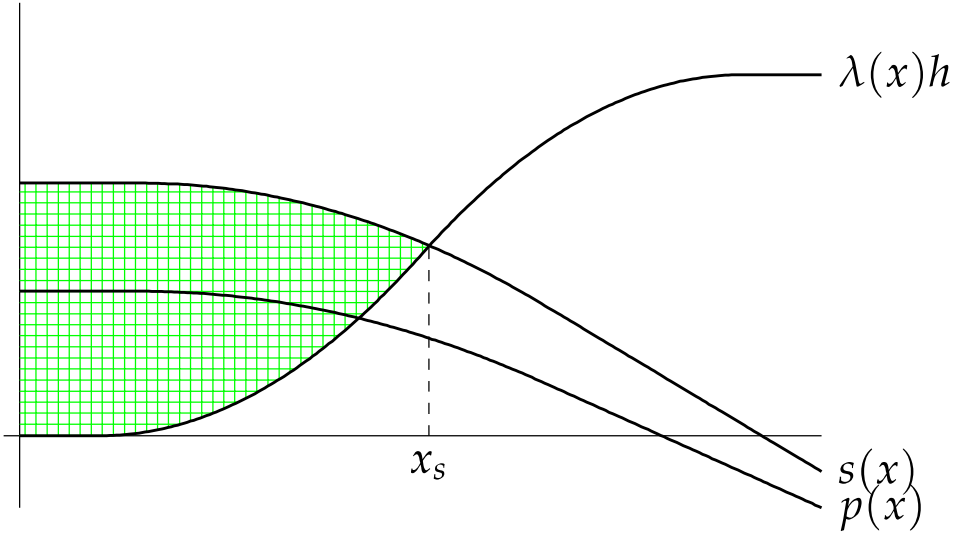


Figure 4 illustrates a socially optimal level of moderation. This is the best moderation that a platform can possibly do without knowing more about which content is harmful and which content is harmless. The green shaded region represents total social welfare under optimal moderation. The platform should host all content to the left of x_s and take down all content to the right of x_s.

Now consider the platform's profits. Since its marginal revenue is $p(x)$, its total profits from setting its moderation level to x^* are the area between $p(x)$ and the x-axis from 0 to x^*. By similar reasoning to the above, the platform maximizes its profits by setting x^* such that

$$p(x^*) = 0.$$

Call this point x_p, the **platform profit-maximizing moderation level**.[44]

44. If there is no such value, which happens when the platform makes positive revenue from all content, the platform should set $x^* = x_{\max}$ and host all content.

Figure 5: Undermoderation under immunity

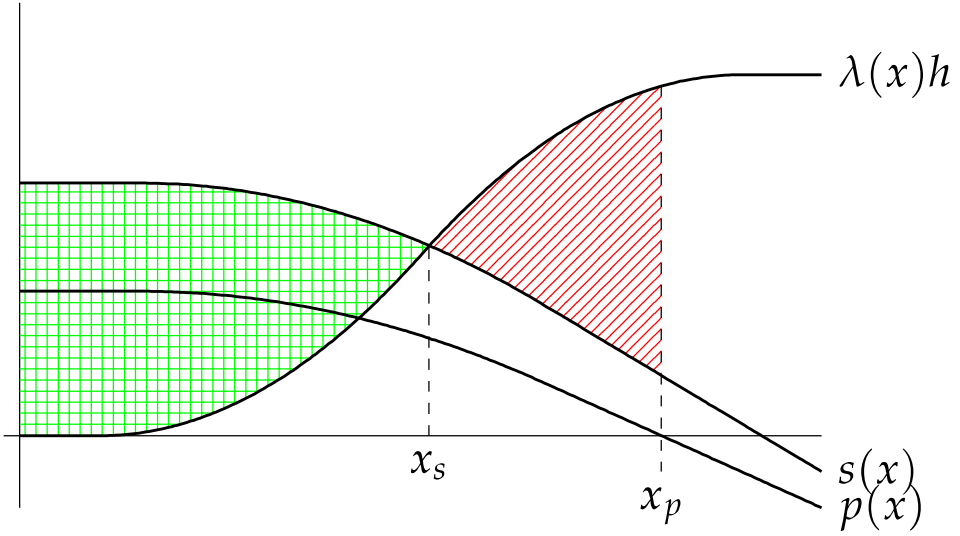
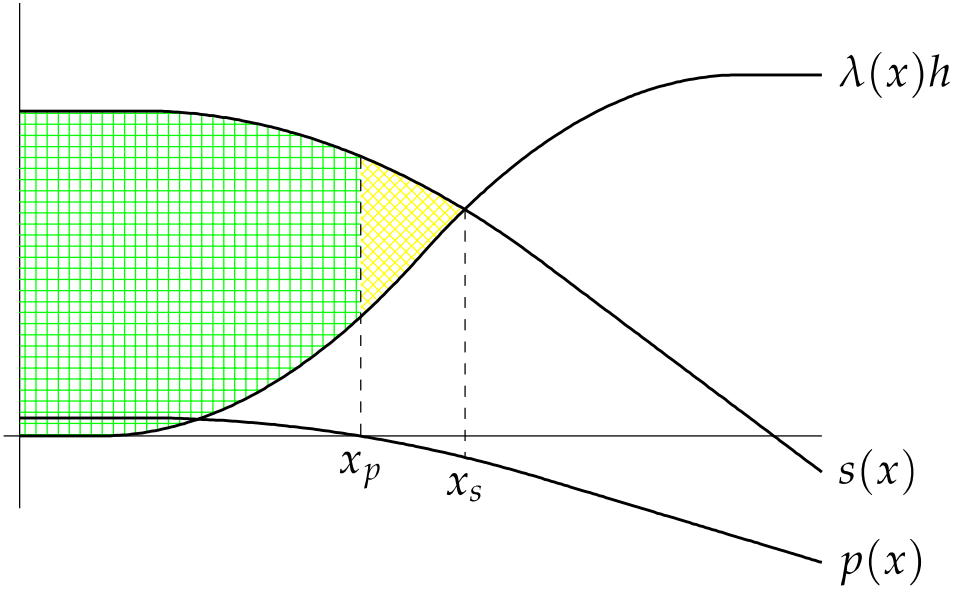


Figure 6: Overmoderation under immunity

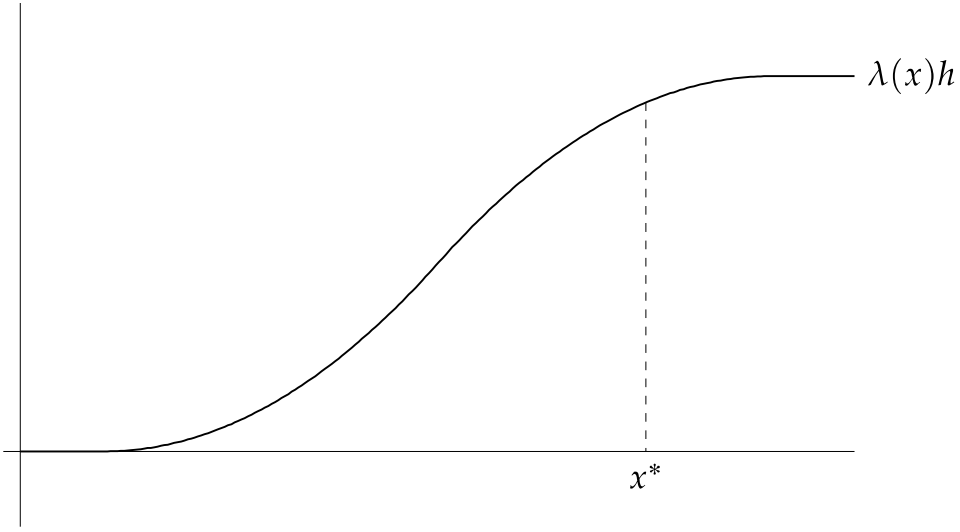


As Figures 5 and 6 show, there is no necessary relationship between x_s and x_p. In Figure 5, the platform undermoderates. It makes money from content that is bad for society, so $x_p > x_s$ and the platform leaves up content that it should ideally take down. The red striped region is the net social loss

from hosting too much content. But in Figure 6, the platform overmoderates. It loses money on content that is good for society, so $x_p < x_s$ and the platform takes down content that it should ideally leave up. The yellow diamond region is the foregone social benefits from content the platform could have hosted but did not.

One way of understanding why both undermoderation and overmoderation are possible is that there are two different effects at work, with opposite signs. On the one hand, the platform fails to internalize the full social benefits of the content that it hosts: $s(x) > p(x)$. On the other hand, when content is harmful the platform does not internalize the harms to third parties: $\lambda(x)h$. On the minimal assumptions we have made, either one of these two effects could dominate. In the real world, both undermoderation and overmoderation are problems that lawmakers have thought serious enough to try to fix. Parts III and IV discuss their various responses in detail.

Figure 7: Blanket immunity



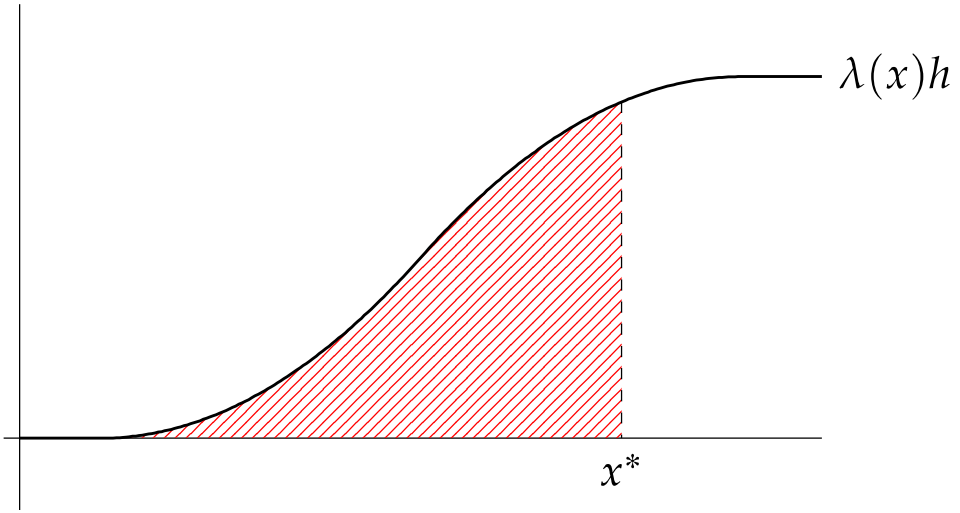
So far, we have been considering a model in which the platform is subject to a legal regime of **blanket immunity**. Figure 7 illustrates this very simple rule. Regardless of where the platform sets x^*, it is not liable for any of the harms that result.

C. STRICT LIABILITY

The essential premise on which any form of liability depends is that some conduct is harmful. The standard law-and-microeconomic response to

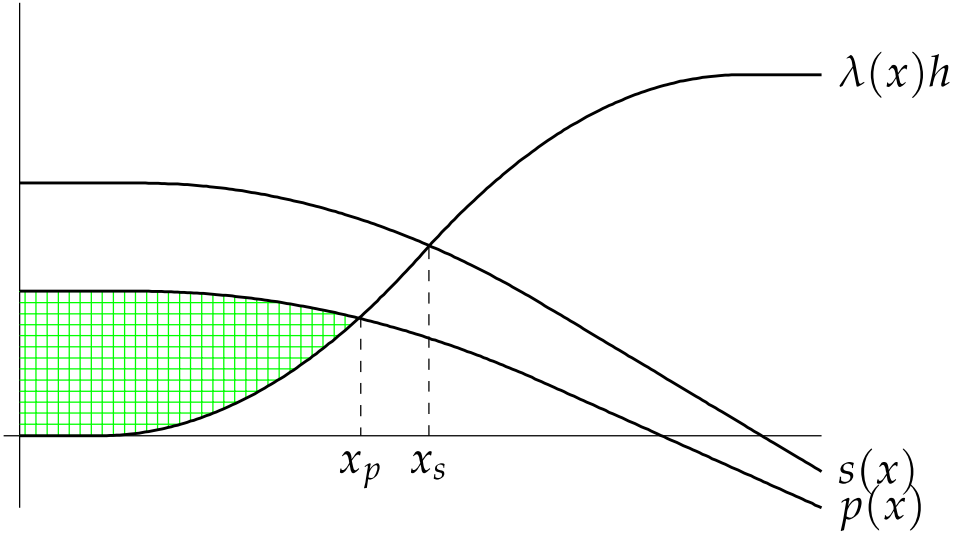
harmful conduct is *strict liability*. If a widget factory is forced to compensate everyone who is injured by defective widgets, the factory will take exactly those manufacturing precautions that are cost-justified. Once the factory internalizes the harms it causes, its incentives are aligned with society's.

Figure 8: Strict liability



For a platform, that conduct is hosting content, and the strict-liability measure of damages is the harm that results from the content that the platform hosts. Figure 8 illustrates that if the platform sets its moderation threshold at x^*, it is liable for all harm caused by the content that it carries (and for none of the harm that would have been caused by content that it could have carried and did not).

Figure 9: Platform's optimal behavior under strict liability



Note that the platform pays no damages for the fraction $1 - \lambda(x)$ of content that is actually harmless. But for the fraction $\lambda(x)$ of content that is harmful, the platform pays total damages of $\lambda(x)h$.

Thus, under strict liability, the platform's marginal profits are $p(x) - \lambda(x)h$. Its profit-maximizing moderation level x^* is defined by

$$p(x^*) = \lambda(x^*)h. \tag{2}$$

Figure 9 illustrates the results. The platform sets its moderation level where its revenue curve $p(x)$ and the damages it must pay $\lambda(x)h$ cross. At that point, its revenues from carrying additional content are exactly cancelled out by the harm that content causes (and hence the damages it must pay).

It follows that *strict liability always results in overmoderation*. Because $p(x) < s(x)$, the platform's profit curve $p(x)$ always intersects the expected-harm curve $\lambda(x)h$ to the left of where the social-benefit curve $s(x)$ intersects $\lambda(x)h$. Thus, $x_p < x_s$.

Figure 10: Strict liability results in overmoderation

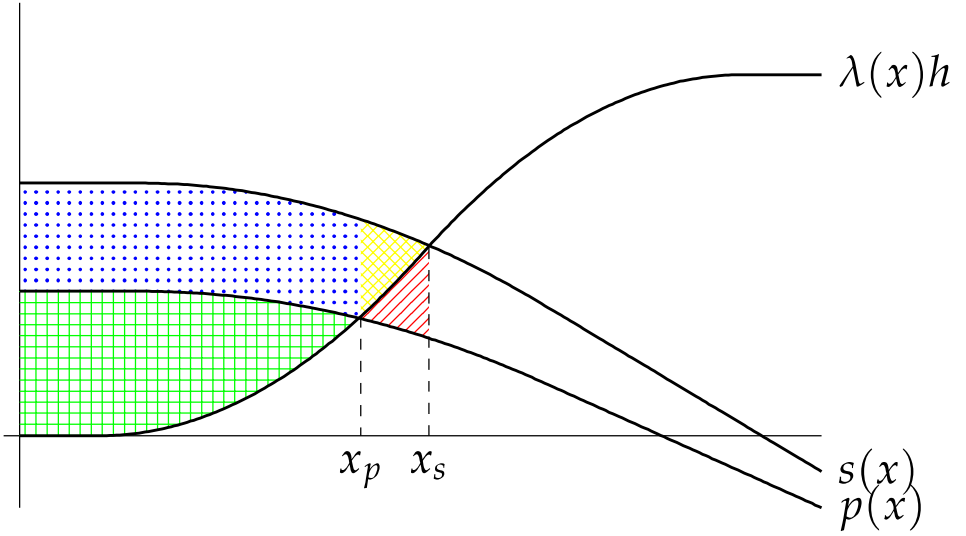


Figure 10 illustrates how strict liability causes overmoderation. The green checked region is the platform's profits, which become 0 exactly at $x_p$where the platform stops hosting content). The blue dotted region is the additional spillover social benefit from the content the platform hosts. Between x_p and x_s, it is unprofitable for the platform to host more content because it would have net losses equal to the area of the red striped region. But content in that range is beneficial overall for society. Society suffers a welfare loss equal to the area of the yellow diamond region from content that platform could have hosted but did not. This content is unprofitable to the platform but beneficial to society, because $p(x) < \lambda(x)h < s(x)$.

D. COSTLESS INVESTIGATIONS

The final moving piece of our model is that a platform can investigate content that it suspects of being harmful. Specifically, we add the option that the platform can pay a cost $c \geq 0$ per unit of content to investigate and determine with certainty whether each item is actually harmful.

To get intuition for how this possibility affects the platform's incentives, we start by presenting extreme cases. When investigation is infeasibly costly, i.e., $c \to \infty$, this model collapses into the previous one because there are no circumstances under which the option to investigate is worth exercising.

On the other hand, when investigation is costless, i.e., $c \to 0$, the platform can perfectly distinguish harmful content and harmless content. That means it

is possible for the platform to take down the harmful content while leaving up the harmless content. From the regulator's perspective, that is exactly what it should do: take down every piece of harmful content and leave up every piece of harmless content.

Naively, it might seem like the effect of costless investigation would be to remove the harm curve $\lambda(x)h$ from the picture, so that the platform earns all the revenue under $p(x)$ and society realizes all the value under $s(x)$. But this is not quite right, because the *harmful content still must be removed*. This means that the platform must forego the revenue, and society the benefits, from the fraction $\lambda(x)$ of content that is removed.

We define $p^*(x) = \big(1 - \lambda(x)\big)p(x)$, i.e., the profits the platform can make by hosting only harmless content. Similarly, we define the corresponding function $s^*(x) = \big(1 - \lambda(x)\big)s(x)$ for social benefits. These new functions represent the maximum revenue and social benefit, respectively, that it is possible to realize with perfect knowledge about which content is harmful.

Figure 11: Platform revenue and social benefits with costless investigations

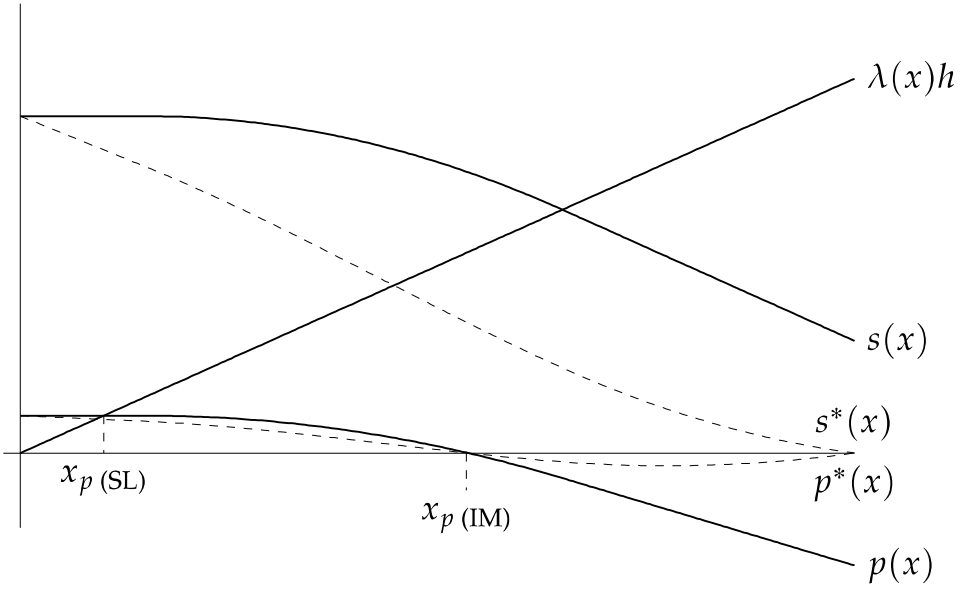


Figure 11 illustrates the platform's profits $p^*(x)$ and social benefits $s^*(x)$ from hosting only harmless content.[45] Their behavior is subtle. $s^*(x)$ starts off equal to $s(x)$ when all content is harmless, and the platform need not remove any content. It immediately dips below $s(x)$ as content must be removed because there is less content available to generate surplus. Eventually, it ends up equal to 0 because all the content is harmful so there is nothing left to host. Similarly, $p^*(x)$ starts off equal to $p(x)$ and immediately dips beneath it. A twist is that $p^*(x)$ becomes 0 exactly when $p(x)$ does because that is the point at which all content, harmful and harmless, is valueless to the platform. From then on $p^*(x) > p(x)$, because the platform saves money by not hosting content that would be unprofitable for it. But like $s(x)$, it eventually ends up equal to 0 because there is nothing left to host.

It is a little difficult to see visually in Figure 11, but costless investigation is always good for society, and society is always better off if the platform removes the harmful content that it knows about. Algebraically, the benefit function with omniscient moderation $s^*(x)$ is always greater than the benefit function with oblivious moderation $s(x) - \lambda(x)h$.[46] Note that society will now prefer to host all harmless content up to the point at which $s(x) = 0$.[47]

From the platform's perspective, omniscient moderation is also never a bad thing. Under immunity, the platform does not care about harmful and harmless content; it still sets its moderation level at $x_{p\,(\mathrm{IM})}$ and it is no worse off. (The "IM" stands for "immunity.") Under strict liability, the platform will still set its moderation level at $x_{p(\mathrm{IM})}$ but it will also remove all of the content $x < x_{p\,(\mathrm{IM})}$ that is actually harmful. This eats into the platform's profits compared with immunity—it makes $p^*(x)$ instead of $p(x)$—but compared with where it would be under strict liability with oblivious moderation it is much better off. Because it does not actually have to pay the harm $\lambda(x)h$, it can move its moderation level from $x_{p\,(\mathrm{SL})}$ to $x_{p\,(\mathrm{IM})}$. (The "SL" stands for "strict liability.") With costless investigation and strict liability, the platform will typically overmoderate for the simple reason that $p(x) < s(x)$, meaning some harmless content may be unprofitable but socially beneficial. The platform may be willing to host more content, but society's preferred moderation level has also shifted to the right.

45. The harm curve $\lambda(x)h$ has been straightened out and the curves $s(x)$ and $p(x)$ separated to make the diagram easier to read.

46. This follows from the definition of $s^*(x)$ and the postulate that $h > s(x)$.

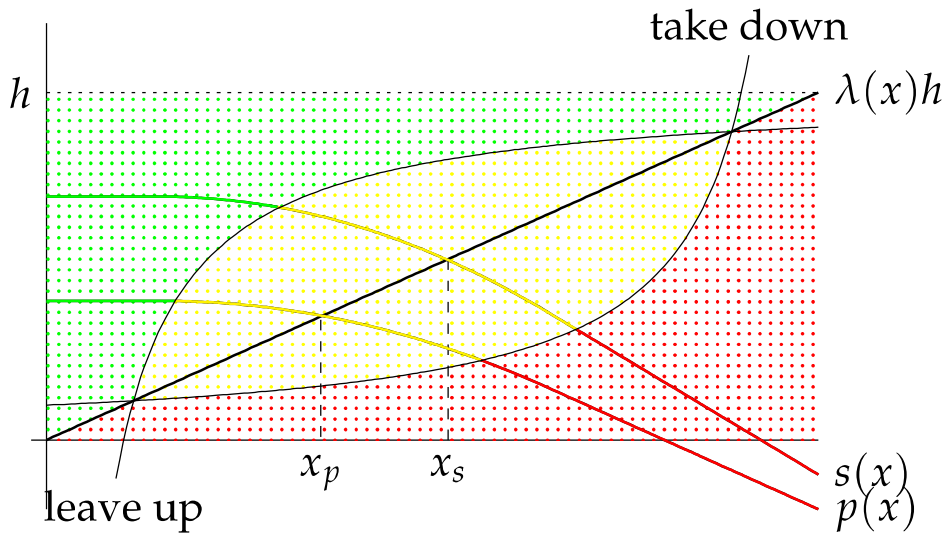
47. Or if there is no such point (as in Figure 11), simply to host all harmless content.

E. COSTLY INVESTIGATIONS

Now consider what happens when an investigation is costly but feasible. Now the platform must make real decisions about whether to investigate because sometimes investigation is worthwhile and sometimes it is not. The platform has three options for any given item of content: leave it up, take it down, or investigate. It is easy to see that the platform will only investigate content where its decision depends on the results of the investigation, i.e., the platform will take the content down if the investigation finds the content harmful and leave it up otherwise. (If the platform intended to take down the content regardless, it could save c by omitting the investigation, and similarly if it intended to leave up the content regardless.)

Thus the expected value to society for content at x is 0 if the platform takes down the content, $s(x) - \lambda(x)h$ per unit if it leaves the content up, and $\big(1 - \lambda(x)\big)s(x) - c$ per unit if it investigates, i.e., the value of a harmless piece of content times the probability that the content is harmless, minus the cost of investigating all content. Intuitively, the platform should prefer takedown for content with $\lambda(x)$ close to 1 and should prefer leaving it up for content with $\lambda(x)$ close to 0, with an interval of investigation somewhere in the middle.

Figure 12: Investigation of intermediate content under strict liability



The regulator is indifferent between takedown and investigation when the value of the content that investigation will allow to remain up minus the costs of investigation ($\big(1-\lambda(x)\big)s(x)-c$) exactly equals the value of taking all content down (which is simply 0). Doing out the math, takedown and investigation are equally efficient when

$$\lambda(x) = 1 - \frac{c}{s(x)}.$$

When c approaches 0, this converges to $\lambda(x) = 1$, i.e., the right end of the investigation interval approaches $x_{\max}$. That is, as the costs of investigation decrease, it is almost always better to investigate than to take down suspected-bad content without first checking.

The regulator is indifferent between investigation and leaving up when the value of the content that investigation will allow to remain up minus the costs of investigation ($\big(1-\lambda(x)\big)s(x)-c$) exactly equals the benefits of all the content minus the costs of the harmful content ($s(x)-\lambda(x)h$). Doing out the math, investigating and leaving up are equally efficient when

$$\lambda(x) = \frac{c}{h - s(x)}.$$

When c approaches 0, this converges to $\lambda(x) = 0$, i.e., the left end of the investigation interval approaches 0. That is, as c decreases, it is almost always better to investigate than to leave up the suspected-good content without first checking.

Put another way, as c decreases, the ideal investigation interval expands to cover more and more content. But as c increases, the investigation interval shrinks and eventually vanishes.[48] When this bound is exceeded, it is never worthwhile from society's perspective for the platform to investigate. It should instead act on the basis of the imperfect information it already has.

These results show that a rational regulator should want platforms to invest resources in investigating only when the cost of investigation is sufficiently low, and then only for a range of intermediate cases where the harmfulness of the content is sufficiently unclear. For content that is highly likely or highly unlikely to be harmful, individual investigation is unnecessary and inefficient. Note that this interval contains x_s—in a sense, affordable

48. It vanishes when:

$$c > \min_{x\in[0,x_{\max}]} s(x)\frac{h-s(x)}{2s(x)-h}.$$

investigations expand the cutoff from a sharp on-off to a range warranting a closer look.

Figure 12 illustrates the intermediate range where investigation is justified.[49] The green dotted region is where no investigation is needed, and the platform should leave up all content; the yellow dotted region is where it should investigate and act accordingly; and the red dotted region is where no investigation is needed and the platform should take down all content. The curve labeled "leave up" is the dividing line between the region where investigation is better than leaving content up and vice versa. The curve labeled "take down" is the dividing line between the region where investigation is better than taking content down, and vice versa. These are two-dimensional regions because whether it is rational to investigate or not depends both on $\lambda(x)$ (the horizontal axis) and on $s(x)$ (the vertical axis). As the probability of content being harmful increases (i.e., as one moves horizontally to the right), one starts in a region where it is optimal to leave content up, passes through a region (possibly zero-width) where investigation is optimal, and then moves into a region where it is optimal to take content down. Similarly, as the value of content increases (i.e., as one moves vertically upwards), the optimal policy changes from takedown to investigation to leaving content up. If the curve $s(x)$ passes through the investigation-justified region at all, then x_s lies within it.

Figure 12 also illustrates the dependence of investigation on c. As c decreases, the upper limit moves upwards and the lower limit moves downwards, increasing the size of the region where investigation is justified. As c increases, these limits converge until eventually the region vanishes entirely. In this case, investigation is never justified, and we are back to the previous model, where $\lambda(x)h$ marks the dividing line between taking down and leaving up.

A nearly identical analysis applies to a platform's incentives under strict liability.[50] Because the platform internalizes all the harm that it causes, the only change is to substitute the platform's private profit $p(x)$ for the overall social value $s(x)$. If there is any range for which investigation is justified, it will contain $x_{p\,(\mathrm{SL})}$. A little algebraic manipulation shows that the platform's preferred interval of investigation is always *shifted left* from the regulator's

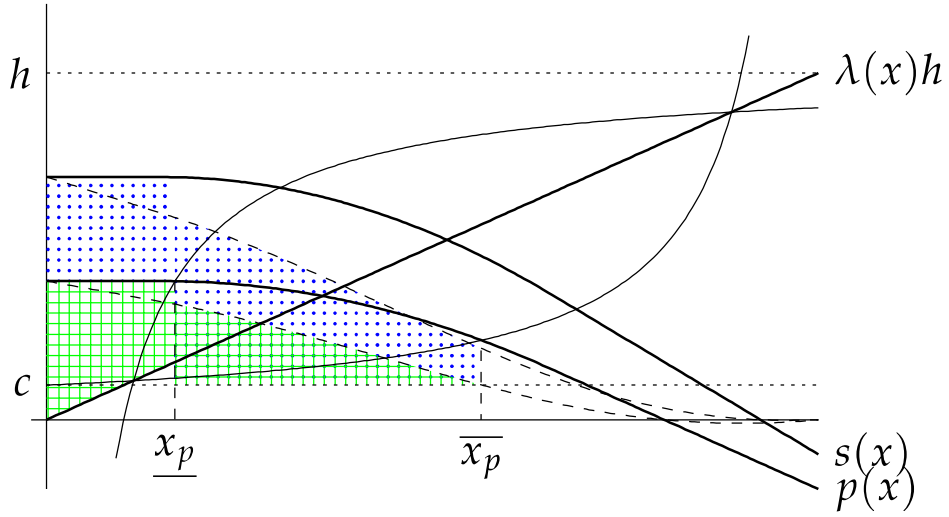
49. Again, for simplicity of illustration, $\lambda(x)$ is shown as a straight line, but the same results hold in the general case where it is any weakly increasing function that goes from 0 to 1 on the interval $[0, x_{\max}]$.

50. Under blanket immunity, a platform will never investigate. Instead, it will always choose to leave all content up.

preferred interval.[51] Intuitively, because the platform has less at stake, it will be more likely to remove content rather than investigating and also more likely to investigate content rather than leaving it up.

Figure 13 shows the platform's profits and social welfare under strict liability when the platform can investigate content. The upper dashed line is $s^*(x)$ and the lower dashed line is $p^*(x)$—respectively, the social benefit and platform-profit effects of omniscient moderation. The difference is that now this moderation must be paid for with investigations, so the net social benefit *if the platform investigates* is $s^*(x) - c$ and the net profit *if the platform investigates* is $p^*(x) - c$.

Figure 13: Platform's profits and social benefits under strict liability with costly investigation



The platform's profits are the green gridded region, and the additional positive externalities for society are the blue dotted region. Beneath the lower limit of investigation, $\underline{x_p}$, matters are as before: the platform's marginal profit is $p(x)$ (income) minus $\lambda(x)h$ (expenses) and marginal social welfare is

51. To be precise, at the lower end

$$\frac{c}{h - s(x)} < \frac{c}{h - p(x)},$$

and at the upper end

$$1 - \frac{c}{p(x)} < 1 - \frac{c}{s(x)}.$$

$s(x) - \lambda(x)h$. When $\lambda(x)$ crosses into the region where investigation is optimal, the platform's marginal revenue is now defined by the difference between $p^*(x)$ (income) and c (expenses). At this point, both income and expenses shift discontinuously downward. The platform is taking in less revenue now that it is removing some content, but that drop is exactly offset by the savings from investigating rather than paying damages. Marginal social welfare discontinuously decreases—intuitively, because society has more to gain from beneficial content and would not have started investigating until later. In this region of investigation, both profits and welfare decrease faster than they did under leave-it-all-up, as more and more content is removed. But this steeper decrease is more than offset by the fact that costs are now constant at c, rather than increasing with $\lambda(x)$. At the upper limit of investigation, $\overline{x_p}$, the platform's marginal profit is zero, so it switches to taking all content down, which zeroes out both marginal profit and marginal welfare going forward. Again, it is visually apparent that the platform is making different tradeoffs than society—it would still be socially beneficial at $\overline{x_p}$ for the platform to continue investigating content.

In short, the ability to investigate increases both social welfare and the platform's profits, but it does not automatically align the platform's incentives with society's incentives.

F. COLLATERAL CENSORSHIP

It is critical to understand why and when strict liability causes overmoderation. Strict liability causes the platform to internalize the harms from the content it carries, but not the offsetting benefits. This asymmetry between harm (for which it faces liability) and benefits (for which it is not compensated) pushes the platform to remove more content than an omniscient regulator would.

This overmoderation fundamentally depends on the platform's imperfect information about content. If the platform could distinguish harmless and harmful content without incurring costs, then strict liability would be efficient. It would be feasible to expect the platform to separate the two and remove only the harmful content. But given imperfect information, the platform *cannot tell with certainty* which content is harmless and creates net positive externalities and which content is harmful and creates net negative externalities. A platform facing strict liability consistently overmoderates and removes more harmless content than it should from society's perspective.

Thus, our model validates Felix Wu's argument for intermediary immunity.[52] The combination of positive externalities and imperfect information causes a platform subject to strict liability to engage in collateral censorship. The platform has less at stake than an original speaker (positive externalities) and responds by removing good content as well as bad (imperfect information). These conditions are jointly necessary and sufficient; if there are no positive externalities (i.e., $s(x) = p(x)$) or the platform has perfect information (i.e., $\lambda(x) = 0$ or $\lambda(x) = 1$ for all content), then strict liability is efficient.

It is worth dwelling for a bit on the nature of these positive externalities. A widget factory might come close to capturing the full social value of the widgets it makes. But a platform does not, for at least two reasons.

First, a platform's "product" is not widgets but speech. Speech consists of information, and information is a public good. Once it has been shared with one listener, then neither the speaker nor the platform can easily prevent them from sharing it with others. A dance video that goes viral on TikTok will be reposted to Twitter and YouTube; the information in a plumbing tutorial will be retained in the minds of viewers and shared with others. All the third-party value is an externality from both the speaker's and the platform's perspectives.[53]

The second source of positive externalities is that platforms do not even capture the full value to speakers of the content they host. As Felix Wu convincingly argues, the value to a user of posting content to a platform is typically much larger than the value to the platform of hosting that content.[54] A platform does not have an original speaker's incentives. This point holds true even for non-speech platforms. For example, Airbnb captures only part of the value that apartment hosts and guests enjoy from rentals made through the platform.[55]

As Wu explains, speech law already provides heightened protections for original speakers—and yet intermediaries have protections that are higher

52. *See* Wu, *supra* note 1.

53. *See* C. Edwin Baker, *Giving the Audience What It Wants*, 58 OHIO ST. L.J. 311 (1997).

54. *See* Wu, *supra* note 1, at 303–8.

55. *See* Chiara Farronato & Andrey Fradkin, *The Welfare Effects of Peer Entry: The Case of Airbnb and the Accommodation Industry*, 112 AM. ECON. REV. 1782, 1783 (2022) (estimating that in 2014 Airbnb generated "$112 million in peer host surplus, or about $26 per room-night"). *See generally* Erik Brynjolfsson, Avinash Collis & Felix Eggers, *Using Massive Online Choice Experiments to Measure Changes in Well-Being*, 116 PROC. NAT'L ACAD. SCI. 7250 (2019) (estimating value to consumers of numerous online platforms).

still. [56] Speakers have private motivations for speaking: financial, self-expression, reputation-building, community-building, or even revenge. Platforms share their speech but not their motivations.

Platforms also differ from speakers in that speakers generally have much better information about the harmfulness of their speech. A speaker knows whether there is a factual basis for allegations of corruption or harassment; a platform does not. A speaker knows whether they wrote a song themselves or copied it from someone else; a platform does not. A speaker is much less likely to be chilled from harmless speech by the threat of liability for harmful speech.

Whether social welfare is higher under strict liability or immunity depends on the parameters of the model: $p(x)$, $s(x)$, h, and $\lambda(x)$. Strict liability always leads to overmoderation; immunity could either undershoot or overshoot the efficient level of moderation. Generally speaking, a blanket immunity regime is most justified when there are large positive externalities (a large difference between $s(x)$ and $p(x)$), highly imperfect information ($\lambda(x)$ has a large intermediate region that is not close to 0 or to 1), and socially harmful content is also unprofitable ($x_p < x_s$). There is a strong argument that these conditions describe many categories of content moderation today.

G. THE MODERATOR'S DILEMMA

Now we are in a position to appreciate the crucial policy arguments at the heart of § 230. Famously, § 230 was enacted against the backdrop of two judicial decisions on the liability of online intermediaries, *Cubby v. CompuServe* and *Stratton Oakmont v. Prodigy*. In *Cubby*, the court held that CompuServe could not be held liable for user-posted content where it "neither knew nor had reason to know" that the content was defamatory.[57] But in *Stratton Oakmont*, the court held that Prodigy could be held liable for user-posted content, even where it lacked such knowledge.[58] Both courts treated the cases as involving imperfect information—the issue was how a platform *without* specific knowledge should be treated.

Notoriously, the *Stratton Oakmont* court distinguished *Cubby* on the grounds that Prodigy's "conscious choice, to gain the benefits of editorial control, has opened it up to a greater liability than CompuServe and other computer networks that make no such choice."[59] On this reasoning, moderated services

56. Wu, *supra* note 1, at 304.

57. Cubby, Inc. v. CompuServe, 776 F. Supp. 135, 141 (S.D.N.Y. 1991).

58. Stratton Oakmont, Inc. v. Prodigy Servs. Corp., No. 031063/94, 1995 WL 323710, at *3 (N.Y. Sup. Ct. May 24, 1995).

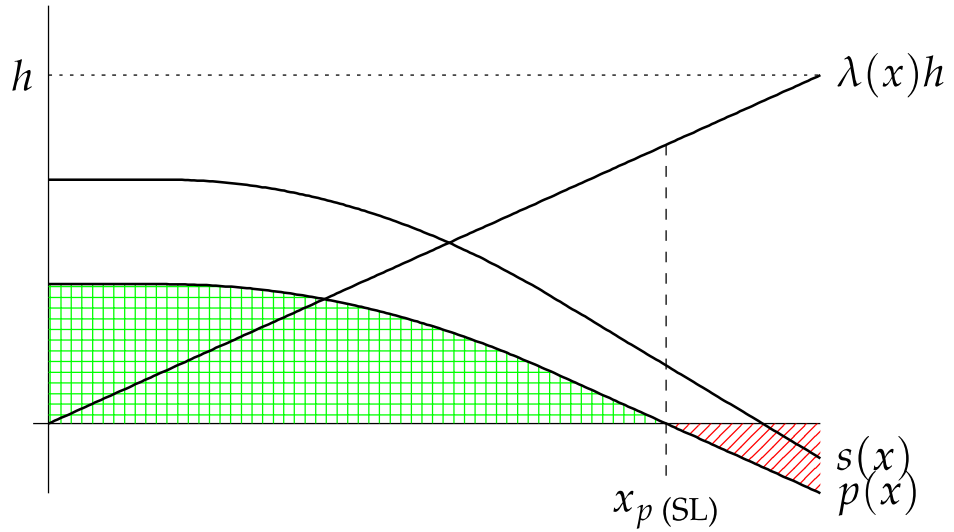
59. *Id.* at *5.

like Prodigy that exercise "editorial control" face strict liability, whereas unmoderated services like CompuServe that exercise no editorial control are immune.

In terms of our model, the rule in *Stratton Oakmont* forces platforms to make a choice. If they host *all* content, they face no liability. But if they remove *any* content, they are strictly liable for the harms caused by any content they do not remove.

Section II.E of this Article analyzed the platform's behavior if it chooses to moderate and thus commits to optimal investigations to maximize its profits in the presence of strict liability. Figure 11 shows the range of content for which the platform will investigate, and Figure 12 shows the platform's profits (green gridded) and additional social welfare (blue dotted) that result.

Figure 14: Platform's profits if it carries all content



Compare that situation with Figure 14, which shows the platform's profits (green gridded) and losses (red striped) if it chooses simply to carry all content. While the platform ends up taking some losses on the spammy, negative-revenue content at the right, it also makes substantial profits on the positive-revenue content at the left—and since the platform no longer has to pay damages, it can pocket all of that revenue without concern for the resulting harms.

Comparing Figure 12 with Figure 14, it is visually clear that the platform is better off not moderating at all. This is contingent on the precise values of the

parameters in the model, especially its profit function $p(x)$. For a different and lower $p(x)$, the platform might lose so much money hosting the worst of the worst content that it would be better off moderating and accepting liability.

These diagrams also reinforce an important point about moderation: *almost all platforms have their own strong incentives to engage in at least some moderation*. The platform here would moderate at $x_{p\,(\mathrm{SL})}$ even in the absence of liability because the worst content is genuinely bad for the platform and its users. Liability is not the only incentive to moderation, and by putting the platform to the choice between voluntary moderation and immunity, the regulator runs the risk that the platform will choose to give up its voluntary moderation efforts.

Figure 15: Social welfare if the platform carries all content

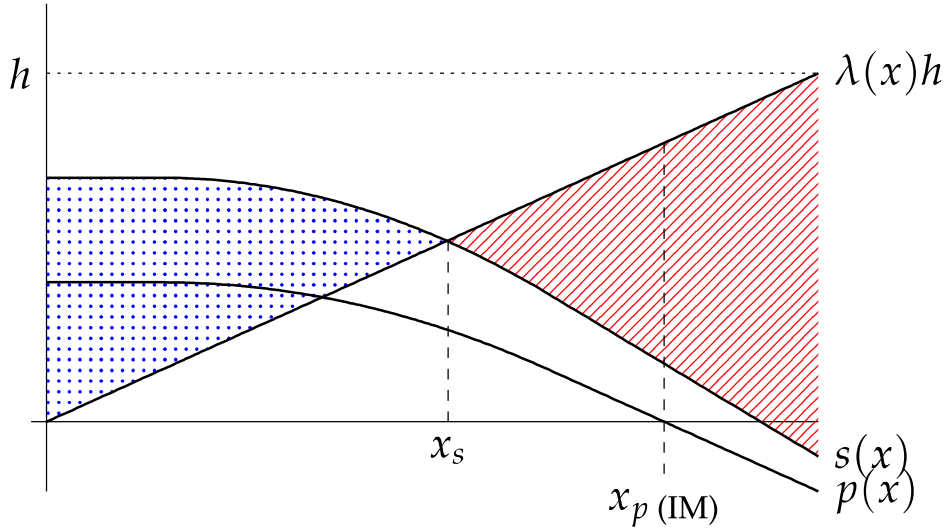


Figure 15 shows the resulting social welfare if the platform chooses to carry all content. The blue dotted region is social benefit and the red striped region is social harm. The red striped area is large. The additional social welfare loss from $x_{p\,(\mathrm{IM})}$ to $x_{\max}$ is particularly substantial. Intuitively, the content that the platform is most selfishly interested in removing is also the content that the regulator most wants it to remove.

Take a moment to let the implications sink in. The platform here is much better off carrying all content; society is much worse off as a result. Strict liability induces the platform to increase its moderation effort from $x_{p\,(\mathrm{IM})}$, where it would moderate in the absence of liability at all. Society gains as a result. But when the platform has the option of not moderating at all—or put

another way, when strict liability is the price it must pay for engaging in moderation—the platform is better off turning off its moderation. It no longer has to pay damages for the content it carries, it no longer has to pay investigatory costs, and it can carry content that would have been unprofitably risky before. These gains are more than enough to outweigh the harms to its platform from the negative-value content on the right. From society's perspective, this is a disaster. In trying to encourage the platform to moderate, the regulator has perversely discouraged it from doing so.

Eric Goldman refers the platform's choice as the "moderator's dilemma."[60] The platform wants to moderate in order to improve its offerings for its users. But when moderation also becomes the legal trigger for liability, the platform must consider whether moderation is still worthwhile.

This is why § 230(c) is titled "Protection for 'Good Samaritan' blocking and screening of offensive material."[61] It was enacted to remove the perverse disincentive to moderation created by the rule of *Cubby*. A platform protected by § 230 is now free to move its moderation off of $x_{\max}$ without fear that it will now open itself to liability and be forced to move much further to the left.

IV. POLICY RESPONSES TO UNDERMODERATION

The fundamental challenge of platform liability law is that content has both harms and benefits to society that the platform does not internalize. A profit-maximizing platform makes its decisions based on how much it can make from hosting content, paying no attention to either positive or negative spillovers. We have seen that under blanket immunity, either of these effects can dominate, so both overmoderation and undermoderation are possible.[62] It is technically possible for these effects to cancel out, so that the platform arrives at an appropriate level of moderation on its own. But there is no particular reason to expect that this would be the case. Instead, a particular platform, hosting a particular type of content, with particular harms and benefits, will typically fall on one side or the other.

This Part gives a comparative analysis of the ways that a regulator could respond to undermoderation; the next Part similarly considers responses to overmoderation. We have already discussed strict liability in detail; this Part considers liability on notice, negligence, and conditional immunity. The point

60. Eric Goldman, *Internet Immunity and the Freedom to Code*, 62 COMMC'N OF THE ACM 22, 22 (2019), https://ssrn.com/abstract=3443976.

61. 47 U.S.C. § 230(c).

62. *See supra* Section III.B.

is not to settle on one or another as optimal, but instead to bring out the intuitions behind each and to get a sense of the conditions they depend on.

A. ACTUAL KNOWLEDGE

At common law, a "distributor" of defamatory speech published by a third party (e.g., a bookstore) was liable "if, but only if, [it] knows or has reason to know of its defamatory character."[63] Section 512(c)(1)(A) removes a platform's immunity as to specific material if it has "actual knowledge that the material . . . is infringing"[64] and the platform does not "act[] expeditiously to remove, or disable access to, the material."[65]

These are examples of *actual knowledge*: a platform is liable for harmful content that it hosts, but only when it has specific knowledge that a particular item is harmful. The intuition behind an actual-knowledge regime is that while it might not be feasible to require a platform to *acquire* the knowledge to show that an item of content is harmful on its own, once the platform *has* such knowledge (from whatever source derived), it is reasonable to expect the platform to take action on it.

In our model, actual knowledge corresponds to cases where the cost of investigation c is 0 as to a particular item of content. As we saw in Section II.D, imposing liability for harmful content when $c = 0$ does not distort the platform's incentives. The platform takes down harmful items where $c = 0$ and it is socially optimal for it to do so. This is a strict improvement over immunity. (The platform leaves up harmless items when $p(x) > 0$, which is not socially optimal, but adding an actual knowledge test to a baseline of immunity does not change matters.)

It is crucial, however, that "actual knowledge" actually means actual knowledge. When investigation is costly because $c > 0$, imposing not-actually "actual knowledge" liability does distort the platform's incentives. In Section III.C below, we analyze the platform's responses to a rule that holds it liable when content has a *high probability* of being harmful, and we observe some of the same potential distorting effects as strict liability.

B. LIABILITY ON NOTICE

If someone else is willing to bear the expense of investigating content, then from the platform's perspective, it receives investigation for free. Put another way, the value of a takedown notice is that it reduces the investigation costs as

63. RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. L. INST. 1977).
64. 17 U.S.C. § 512(c)(1)(A)(i).
65. *Id.* § 512(c)(1)(A)(iii).

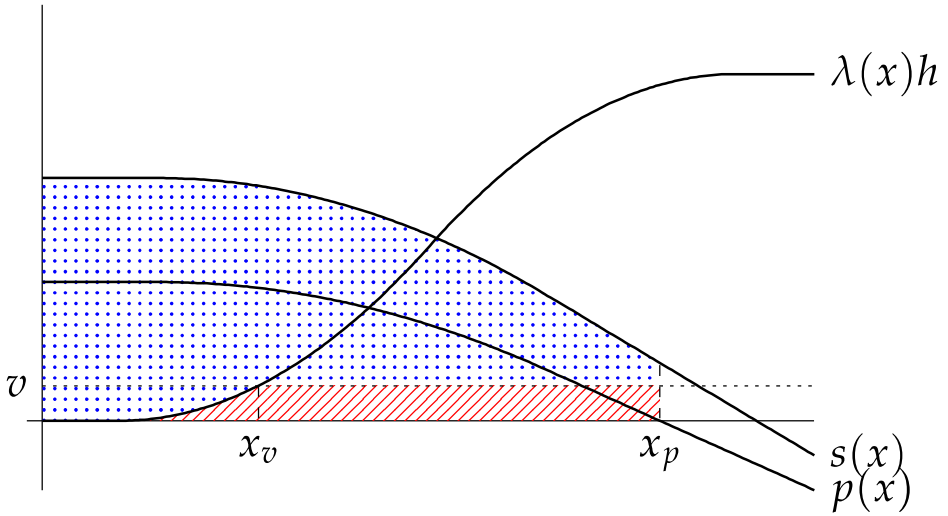
to specific content by narrowing the issues that the platform must investigate. When investigation is expensive, as we have seen, a rational platform will not bother searching for the needle. Instead, it will overmoderate and throw out the entire haystack. But when someone points to an alleged needle, it is far easier for the platform to decide whether it is actually a needle.

The most straightforward way to model liability on notice in our framework is to introduce additional agents: the *victims* of harm, who can investigate content and provide notice to the platform. In this modification, each individual item of content is indexed to a distinct victim, and that victim is the specific person one who suffers the harm if the item is harmful and the platform carries it. The victims, like the platform, can investigate content. Their cost to investigate need not be the same, so we write c_p for the platform's cost of investigation and c_v for the victim's cost. The victims are also able to send notices to the platform for any content they choose, and the platform is liable to the relevant victim for any harm that victim suffers from content about which the platform has received a notice.[66]

Intuitively, it seems that liability on notice should induce the state of affairs depicted in Figure 16. In this figure, the red striped region shows victims' uncompensated harms and investigation costs. The blue dotted region above it shows the social surplus. For content at x, the relevant victim has the option of doing nothing and suffering harm $\lambda(x)h$ or of investigating at cost c_v and giving the platform notice if the content is harmful. For low $\lambda(x)$ they prefer to suffer the harm; for high $\lambda(x)$ they prefer to investigate, with crossover at the point x_v for which $\lambda(x_v)h = c_v$. The platform will always remove any harmful content for which it receives a notice, because a costless removal is better than paying to compensate a harm h that outweighs its profits $p(x)$. Thus, the platform never actually has to pay compensation. (The platform cuts off hosting content entirely all at x^*, where its revenues go negative.)

66. All parties can observe the functions $p(x)$, $s(x)$, and $\lambda(x)$, and the parameters h, c_p, and c_v.

Figure 16: Naive model of liability on notice



But this reasoning is incomplete. The problem is that victims *are not restricted to sending notices for content that is actually harmful*. The victims have a third option for content besides ignoring it and investigating it—they can also send a notice without investigation. In economic terms, liability on notice creates a signaling game between victims and platform. For each item of content, the relevant victim chooses whether to ignore it, investigate and give notice if the content is harmful, or give notice without investigation. The platform either does or does not receive a notice and then chooses whether to take the content down, investigate it, or leave it up. The above reasoning applies only if the signals are all truthful.

When the signals are not truthful, notice on takedown might collapse into strict liability. The victim never investigates but always sends a notice. Because the platform receives a notice regardless of whether the content is harmful or not, the notices are of no use to the platform in distinguishing harmful from harmless content. At the same time, the platform is now legally on notice of all harmful content, and thus subject to strict liability for failure to remove it. The platform faces exactly the same incentives, with exactly the same knowledge, and exactly the same options as in the strict liability case. The notices do no useful work.

This analysis bears out academic criticism of the § 512(c) notice-and-takedown regime. Copyright claimants frequently send notices based on no or minimal investigation, including on content that involves no reuse of copyrightable expression or is obviously a fair use.

One way to make the signal provided by a notice more credible is to make it more expensive to send notices against harmless material. Section 512(f) tries to do this by imposing liability on anyone who "knowingly materially misrepresents" that material is infringing in a takedown notice.[67] Unfortunately, judicial interpretations have almost completely defanged this remedy. Courts have held that a subjective belief of infringement, however unreasonable, is a sufficient defense to a § 512(f) suit.[68] They have also held that even the most cursory investigative process is sufficient.[69] These holdings undermine the effectiveness of notices as signals.

Another way to make a notice more useful is to require it to contain specific evidence of harmfulness, thereby making the platform's own investigation cheaper. To rephrase the standard test for copyright infringement slightly, a claim of copyright infringement requires proof that (1) particular material (2) uses a copyrighted work (3) in a way that infringes.[70] In the abstract, investigation is expensive because a platform must investigate all of its content, compare that content to all copyrighted works, and consider all possible justifications (such as licenses, fair use, etc.). The statutory template for a takedown notice addresses these elements by requiring, respectively, "[i]dentification of the material that is claimed to be infringing . . . and information reasonably sufficient to permit the service provider to locate the material,"[71] "[i]dentification of the copyrighted work claimed to have been infringed,"[72] and "[a] statement that the complaining party has a good faith belief that use of the material . . . is not authorized by the copyright owner, its agent, or the law."[73]

Experience has shown that these three requirements stand on somewhat different footings. Courts have generally been unwilling to relax the requirement of identification of specific material, recognizing that without that specific identification the platform must investigate a vast array of content.[74] And plaintiffs have also been held to the requirement that they identify the

67. *See* 17 U.S.C. § 512(f).

68. *See* Rossi v. Motion Picture Ass'n of Am., 391 F.3d 1000, 1004–05 (9th Cir. 2004).

69. *See* Lenz v. Universal Music Corp., 815 F.3d 1145, 1154 (9th Cir. 2015).

70. Feist Publ'n, Inc. v. Rural Tel. Serv. Co., Inc., 499 U.S. 340, 361 (1991) ("To establish infringement, two elements must be proven: (1) ownership of a valid copyright, and (2) copying of constituent elements of the work that are original.").

71. 17 U.S.C. § 512(c)(3)(A)(iii).

72. *Id.* § 512(c)(3)(A)(ii).

73. *Id.* § 512(c)(3)(A)(v).

74. *See* Perfect 10, Inc. v. CCBill LLC, 340 F. Supp. 2d 1077, 1099–101 (N.D. Cal. 2004), *aff'd in relevant part*, 488 F.3d 1102 (9th Cir. 2007).

relevant copyrighted works.[75] (Indeed, in a world where copyright subsists on fixation, almost every upload will contain material that is copyrighted by someone, so that all of the important questions about the copyright itself go to whether the uploader had the right to do so.) But, as noted above, courts have held that the "good faith belief" required by § 512(c)(3)(A)(v) can be satisfied by a subjective belief, regardless of whether that belief is reasonable or not. Even if the notice-sender acts in bad faith, the damages against them are likely to be nominal at best.[76]

This analysis also shows why commentators have generally regarded liability on notice as producing similar chilling effects to strict liability, even outside the copyright space.[77] It is simply too easy to send a notice against content that is not actually harmful. Proposals to instate some kind of liability on notice need to affirmatively demonstrate that the notices they allow will be credible signals.

C. NEGLIGENCE

Strict liability is not the only form of liability. Another version, which is modeled on the negligence tort, sets an objective standard of care. If the actor complies with the standard of care, it is not liable, even if harm results. But if the actor's conduct falls beneath the standard of care, it is liable for any resulting harm. Although scholars dispute the extent to which the standard of care in negligence is defined mathematically, it is sometimes described in terms of the "Hand formula," $B = P \times L$. Under this formula, an actor is liable for failure to invest in a precaution that would have prevented a harm if the cost of the precaution B is less than the ex ante probability of harm P times the magnitude of the harm L.

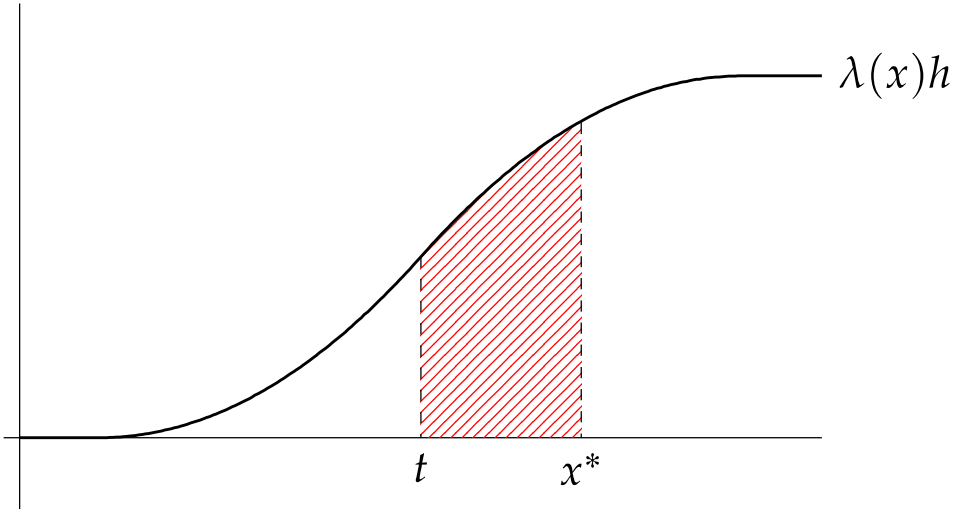
In our model, the regulator imposes negligence liability on the platform by setting a threshold t. The platform is liable for the full harm resulting from hosting any content with $x > t$ but it is not liable for any harm from content with $x \leq t$. Figure 17 illustrates this concept. Note the sharp discontinuity at t.

75. *See id.*

76. *See* Rossi v. Motion Picture Ass'n of Am., 398 F.3d 1000, 1004–05 (9th Cir. 2004); Lenz v. Universal Music Corp., 815 F.3d 1145, 1154, 1156 (9th Cir. 2015).

77. *E.g.*, Wu, *supra* note 1; Schruers, *supra* note 7.

Figure 17: Negligence



The platform's behavior under a negligence regime is identical to its behavior under strict liability, except that it always chooses to leave content up for $x < t$. Thus, the regulator should set t equal to the value of x for which the social benefit of leaving content up is equal to the social benefit of investigation. But this is just the socially optimal lower limit of investigation $\underline{x_s}$. Setting t higher means that the platform will leave up content the regulator would prefer it to investigate (or even take down); setting t lower means that the platform will investigate content the regulator would prefer it to leave up without investigation.

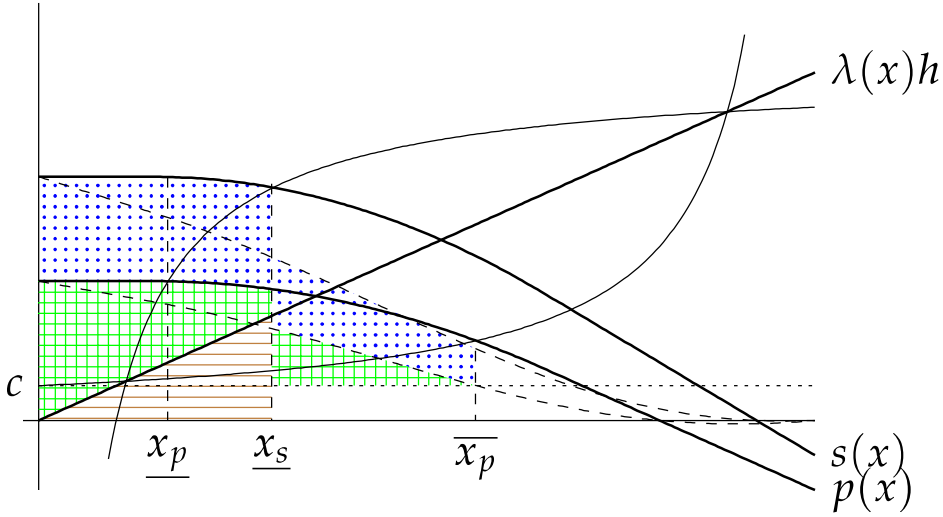
Figure 18 shows the consequences of a negligence rule. Most importantly, it pushes the platform's lower limit of investigation up from $\underline{x_p}$ (where $p(x)$ intersects the lower-limit curve) to $\underline{x_s}$ (where $s(x)$ intersects the lower-limit curve). This is welfare-improving because throughout this range, the value of leaving up ($s(x) - \lambda(x)h$) exceeds the value of investigation ($\big(1 - \lambda(x)\big)s(x) - c$). (Compare this figure to Figure 14, which shows what happens under strict liability.)

There are also distributional consequences. The brown striped region represents uncompensated harm to victims—this region is not part of the social surplus from content on the platform. But it is part of the platform's profits. Note that this is harm that is socially optimal not to attempt to prevent because imposing liability on the platform causes it to inefficiently spend resources investigating. This is not a case where the platform takes down harmless content in ignorance of its harmless nature (that occurs in at higher

values of x, at the right of the diagram, beyond the upper-limit curve). It is a case where the platform *spends too much* on investigation under strict liability, and society is better off overall moving the threshold of liability upwards from 0 (strict liability) to $\underline{x_s}$ (optimal standard of care under negligence).

This system is still not efficient. It gets the platform's incentives right at the boundary between leaving up and investigating, but not at the boundary between investigating and taking down. The platform will still spend too little on investigation at that boundary (from the regulator's perspective) and take down harmless content because it is too similar to possibly harmful content. The optimal negligence rule still results in overmoderation.

Figure 18: Social welfare under negligence



There is an additional challenge. A negligence regime improves on strict liability if the regulator can calculate $s(x)$, h, $\lambda(x)$, and c to set the appropriate threshold. This is not necessarily an easy task, because it involves weighing the full benefits and harms of content, the ex ante likelihood that given content is harmful, and the cost of investigation to make sure. If the regulator sets t too low, it blends into strict liability. If the regulator sets t too high, it blends into immunity. Negligence is always at least as good as one of these two, but it is not necessarily any better.

An example of a negligence rule in platform law is § 512(c)(1)(A)(ii), which removes the platform's immunity as to specific content if it is "aware of facts or circumstances from which infringing activity is apparent" and fails to

remove the content.[78] This exception, known in the caselaw and scholarship as the "red flag" provision, is best understood as a judgment that in certain cases, the probability of infringement is high enough to justify removal.[79] In other words, the red flag provision is a negligence-style rule—beyond some threshold t of high likelihood that content is infringing, the platform will be liable for all such infringing content. Caselaw confirms that t is high.[80] It is not enough that the platform is aware in general that some content is infringing; it must be awareness of "facts that would have made the specific infringement 'objectively' obvious to a reasonable person."[81]

D. CONDITIONAL IMMUNITY

A hybrid of strict liability and immunity is *conditional immunity*. Informally, the platform is immune provided that it keeps total harm small enough. Formally, the regulator sets a harm threshold T. If the total harm caused by the content that the platform hosts is less than or equal to T, the platform's liability is zero (Figure 19). In Figure 19, the yellow dotted region shows harm caused to victims for which the platform is not liable. But if the total harm exceeds this threshold, the factory loses its immunity and is liable for *all* the harm it caused, even those beneath the threshold (Figure 20). In Figure 20, the red striped region shows harm caused to victims for which the platform is now liable. The region that was yellow in Figure 19 is now red in Figure 20. The platform's liability to victims of harm depends on how much other harmful content it allows.

78. 17 U.S.C. § 512(c)(3)(A)(ii).

79. *See, e.g.*, Edward Lee, *Decoding the DMCA Safe Harbors*, 32 COLUM. J.L. & ARTS 233, 251–59 (2008); Viacom Int'l, Inc. v. YouTube, Inc., 676 F.3d 19, 31–32 (2d Cir. 2012).

80. *Viacom Int'l, Inc.*, 676 F.3d at 31–32.

81. *Id.* at 31.

Figure 19: Conditional immunity (below threshold)

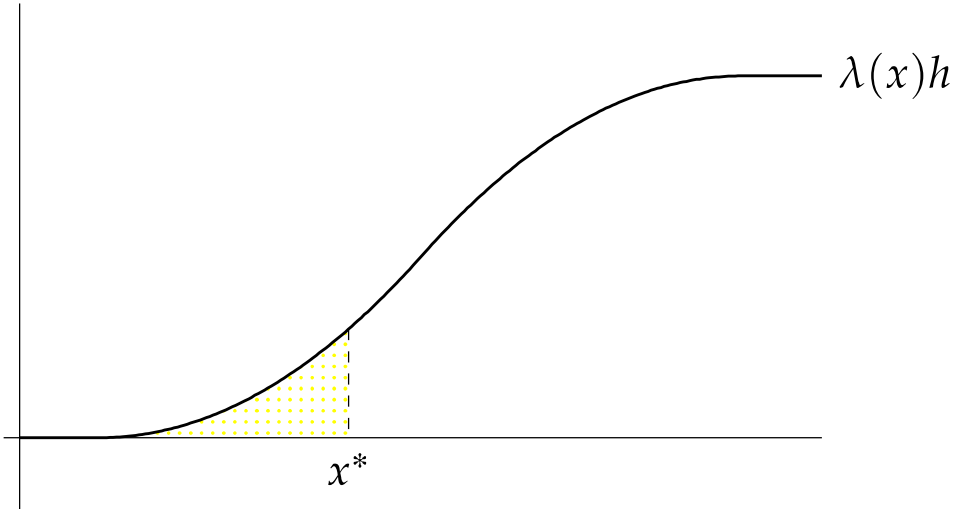
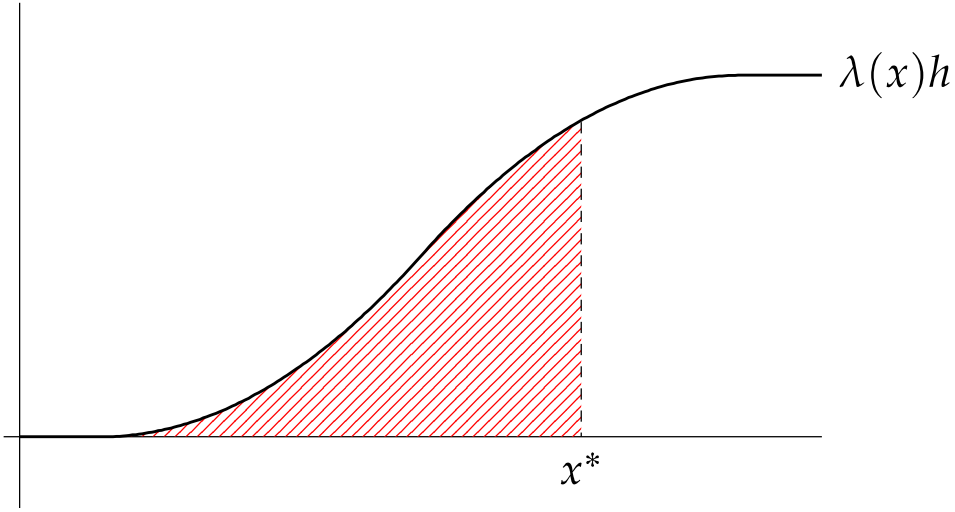


Figure 20: Conditional immunity (above threshold)



An example of conditional immunity is the repeat-infringer provision (RIP) of § 512. To be eligible for the safe harbor at all, a platform must "adopt[] and reasonably implement[] . . . a policy that provides for the termination in appropriate circumstances of . . . repeat infringers."[82] If a platform doesn't do a good enough job at removing content posted by repeat

82. 17 U.S.C. § 512(i)(1)(A).

infringers, it is not eligible for the safe harbor at all, even for material posted by others. Another example of conditional immunity is Danielle Citron and Benjamin Wittes's proposal to condition § 230 immunity on the platform making reasonable efforts to prevent the posting of illegal content.[83]

Both negligence and conditional immunity use a threshold to shape the platform's liability. But they do so in different ways. Negligence imposes liability for *specific content* that exceeds the threshold. Conditional immunity imposes liability for *all content* if total harm exceeds the threshold.

Despite this difference, conditional immunity and negligence have similar incentive effects. Under conditional immunity, the platform in effect has a "budget" of harm it can cause without incurring liability. If the platform has the choice of two items of content at x_1 and x_2 to leave up rather than investigate, where $x_1 < x_2$, it is always better off picking x_1 because $\lambda(x_1)h \leq \lambda(x_2)h$ (i.e., x_1 uses less of the harm budget) and $p(x_1) \geq p(x_2)$ (i.e., x_1 makes more profit for the platform). A similar argument shows that if the platform is choosing between two items to investigate rather than take down, it is always better off choosing the one to the left to investigate. And finally, the platform is best off spending all its budget—it should leave up content until the total harm equals T and then use investigation and takedown to ensure that no further harm ensues.

It follows, therefore, that the optimal level threshold level is

$$T = \int_0^{\underline{x_s}} \lambda(x)h\,dx.$$

If the regulator does so, the platform's behavior and social welfare are exactly the same as under negligence.

There are, however, meta-level concerns about conditional immunity. The first is that the calculation problem is more difficult. The regulator must be able evaluate $\lambda(x)$ and $p(x)$ at every point in $[0, \underline{x}]$, not just at $\underline{x_s}$. A second is that conditional immunity is much more sensitive to errors in this calculation process. Small errors in setting the negligence threshold lead to small changes in the platform's liability. But small errors in setting the conditional immunity threshold can lead to large changes in liability if a platform that thought it qualified for the immunity discovers it did not. The ISP Cox, for example, faced a $1 billion damage award after the court held that its repeat-infringer

83. Citron & Wittes, *supra* note 18, at 455–56.

policy was insufficient to qualify for the § 512(a) safe harbor.[84] Where a negligence regime acts as a price, a conditional immunity has characteristics of a sanction. Prices are more appropriate when the harm can be quantified but the appropriate level of activity is uncertain.[85]

V. POLICY RESPONSES TO OVERMODERATION

A. SUBSIDIES

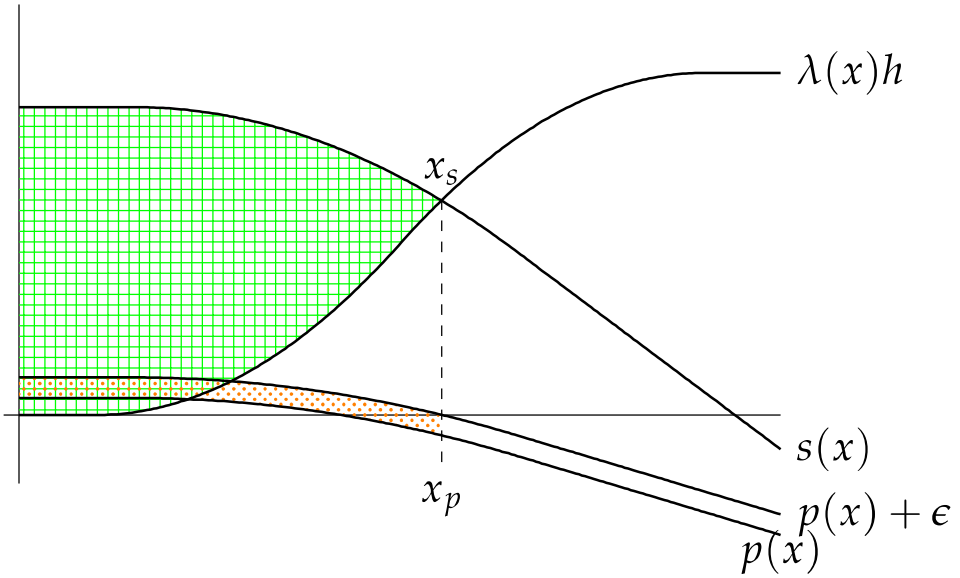
Many responses to overmoderation are familiar from telecommunications and intellectual-property law. One of the most common is *subsidies*, in which the government pays the platform to carry content. Figure 21 shows a case in which the government gives the platform a subsidy of ϵ for any content that it carries. Here, ϵ has been chosen so that it pushes the platform's profits up to the point that $x_p = x_s$ and it carries the socially optimal level of content.

There are at least three challenges in providing subsidies. First, the regulator must accurately estimate x_s, which requires an understanding both the value of content $s(x)$ and the harm of the content $\lambda(x)h$. Second, the regulator must choose an appropriate subsidy ϵ, which requires an understanding of the platform's revenues $p(x)$. And third, the subsidy must be one that the regulator is willing to pay. The orange dotted region in Figure 21 is money that must come from somewhere. It is not a welfare loss to society, just a wealth transfer (ignoring administrative costs and the distortionary effects of taxation, that is). Below-cost mail service is an example of this type of subsidy.

84. Sony Music Ent. v. Cox Commc'ns, Inc., 464 F. Supp. 3d 795, 837–39 (E.D. Va. 2020) (damage award); BMG Rights Mgmt. v. Cox Commc'ns, Inc. 881 F.3d 293, 301–05 (4th Cir. 2018) (safe harbor).

85. *See generally* Robert Cooter, *Prices and Sanctions*, 84 COLUM. L. REV. 1523 (1984).

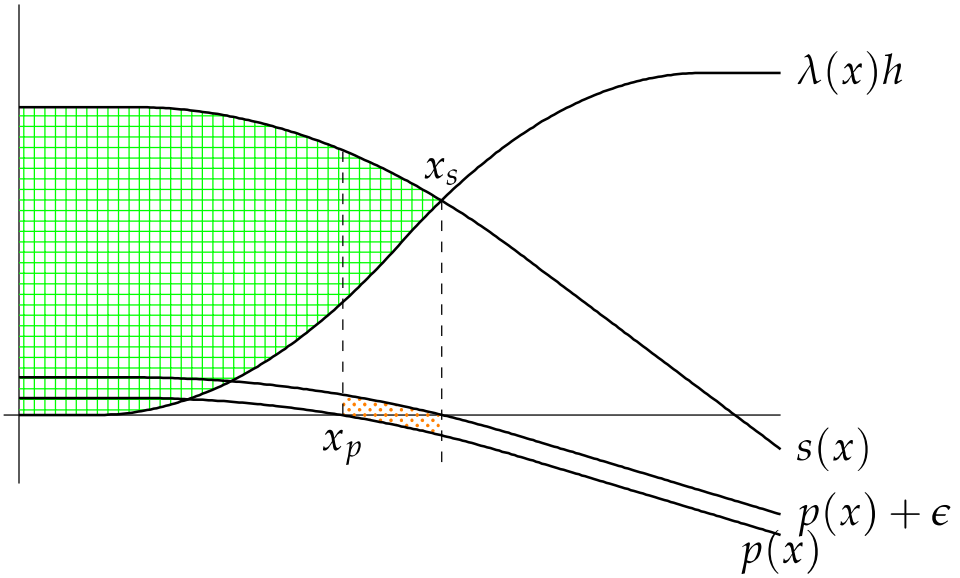
Figure 21: Flat subsidies



A partial solution to the third problem is *targeted subsidies*. Here, the government subsidizes content only in the range where subsidies make a difference in the platform's decision of whether to carry it (between x_p and x_s). This reduces the size of the subsidies required, but it increases the difficulty of the regulatory problem because now the regulator must be able to accurately estimate x^* and not just know the behavior of $p(x)$ in the neighborhood of x_s. For example, the FCC's Universal Service Fund is a targeted subsidy. It helps make broadband internet access more widely available by supporting its availability to people and communities for whom it would not otherwise be profitable for telecom companies to provide it.[86]

86. 47 C.F.R. pt. 54 (2022).

Figure 22: Targeted subsidies



Subsidies can also be provided indirectly, by subsidizing the users who create content and distribute it through platforms and the consumers who receive it. The idea here is that if distribution is more valuable to creators and consumers, they will be willing to pay more to distributors, thus shifting the $p(x)$ curve upwards. There is an argument that the copyright system has some of these features, although it is not typically described in these terms.

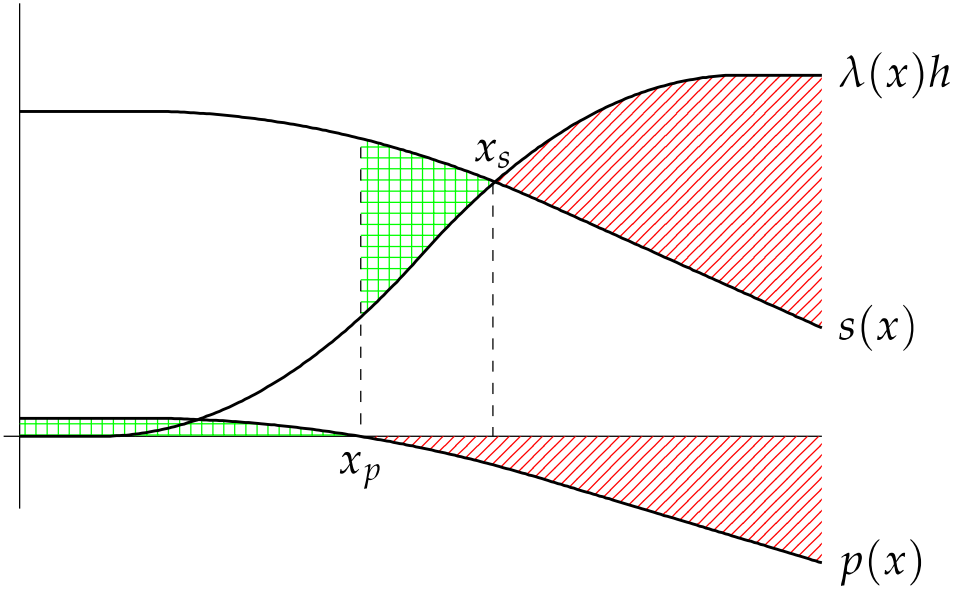
B. MUST-CARRY

Another response to overmoderation is to impose a *must-carry* rule, in which the platform must host all content submitted to it. Formally, the regulator forces the platform to set $x^* = x_{\max}$, i.e., the far right of the diagram. Compared to subsidies, a must-carry system is simpler to design and almost by definition requires less outlay. It also removes discretion from the platform, which may be a concern if the platform has a conflict of interest due to other business lines or does not agree with the regulator's understanding of which content is valuable. Something like this, for example, is a commonly advanced argument for network neutrality.

A must-carry rule, however, must satisfy two conditions to be justified compared with the baseline. First, it must actually result in hosting more worthwhile than worthless content. In Figure 23, the upper green gridded region is the positive-value content that must-carry causes to be hosted, and the upper red striped region is the negative-value content it also causes to be

hosted. If the red region is larger than the green one, must-carry is counter-productive because the bad additional content outweighs the good.[87]

Figure 23: Must-carry



A little more subtly, must-carry can also counter-productively drive a platform out of the market. In Figure 23, the lower green gridded region is the platform's profits from hosting the content it wants to, and the lower red striped region it is losses from hosting the content it is forced to. If the red region is larger than the green one, it is unprofitable for the platform to operate at all, and the platform will rationally shut down rather than comply with a must-carry mandate.

C. LAWFUL MUST-CARRY

An issue with a pure must-carry regime is that it compels platforms to carry content that society itself considers harmful, even illegal. So it is common to see must-carry mandates limited to "lawful" content. The FCC's Obama-era network neutrality regulations had such a carveout,[88] as do the Texas and

87. This analysis omits the investigation option because it is never rational for a platform to investigate content that it is just going to leave up anyway.

88. Safeguarding and Securing the Open Internet, 88 Fed. Reg. 76048, 76096 (Nov. 3, 2023) (to be codified at 47 C.F.R. § 8.2(b)) (prohibiting broadband providers from "block[ing] *lawful* content, applications, services, or non-harmful devices" (emphasis added)).

Florida social media must-carry bills whose constitutionality is currently being litigated.[89]

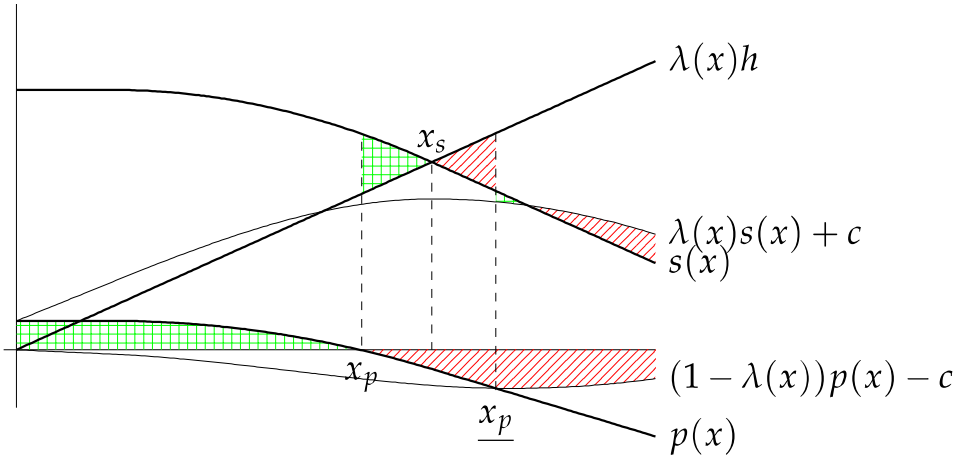
We can model a *lawful must-carry* rule by stating that the platform *must* host all harmless content, but it has discretion whether or not to host harmful content. Of course, to know with certainty whether content is harmless, the platform must investigate it. Thus, under lawful must-carry, the platform has two choices: it can either leave the content up without investigation, or it can investigate it and take it down if harmful.

As Figure 24 illustrates, the platform's marginal revenue from leaving up is $p(x)$, and its marginal revenue from investigation is $\big(1-\lambda(x)\big)p(x)-c$. Thus, the platform finds the two equal when $p(x)=-c/\lambda(x)$, which can only occur when the platform's profit $p(x)$ has gone negative. If these two curves meet at all, we call this intersection $\underline{x_p}$ because this is the point at which the platform starts investigating in the hopes of being able to find and remove harmful unprofitable content. To the left of $\underline{x_p}$, the platform leaves up content, so its profits and social welfare are as above in Section B. But to the right of $\underline{x_p}$, the platform investigates all content and takes down all harmful content. Compared with a flat must-carry requirement, the platform can reduce its losses from the content it is compelled to carry and thus may be better able to keep operating in the face of a lawful must-carry requirement.

Lawful must-carry can also be better for social welfare because the platform will filter out some content that is both harmful and unprofitable. Figure 24 shows that the welfare effects can be subtle and complex. $\underline{x_p}$ creates a discontinuity. To the left, social welfare is the benefits of all content $s(x)$ minus the harm of all content $\lambda(x)h$. To the right, investigation eliminates the harm $\lambda(x)h$ but introduces two new costs: the cost of foregone benefits from removed harmful content $\lambda(x)s(x)$ and the costs of investigation c.

89. *See* NetChoice, LLC v. Att'y Gen., 34 F.4th 1196 (11th Cir. 2022); NetChoice, LLC v. Paxton, 49 F.4th 439 (5th Cir. 2022).

Figure 24: Lawful must-carry



VI. EXISTING AND PROPOSED LAWS

A. SECTION 230

Section 230, in our model, is a blanket immunity regime. The platform is not liable for any harmful content, regardless of its knowledge and regardless of whether it has made any effort to investigate.[90] As discussed in Section II.G, such a regime is a reasonable response to the perverse incentives of *Stratton Oakmont, Inc. v. Prodigy Services Co.*[91] Platforms have their own commercial reasons to moderate content, so it is important not to create a system in which they are disincentivized from moderating at all.

Our model also illustrates the wide range of proposed reforms to § 230. These reforms have profoundly different economic consequences.

To begin, the Citron-Wittes proposal is a straightforward conditional immunity.[92] Courts would be asked to assess a platform's overall moderation efforts and to deny platforms the § 230 safe harbor if they fell beneath that threshold. It therefore functions like the RIP limitation on § 230 and can be expected to have some of the same consequences, including the occasional massive verdict against a platform that miscalculates the required level of effort and a corresponding *in terrorem* effect against other platforms that will cause

90. 47 U.S.C. § 230.

91. *See* Stratton Oakmont, Inc. v. Prodigy Servs. Corp., No. 031063/94, 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995) (holding a platform liable for user-posted content even where it lacked knowledge of the content).

92. *See* Citron & Wittes, *supra* note 18.

them to engage in overmoderation due to the uncertainty they face about their legal exposure.

Other scholars have proposed that *Zeran v. America Online, Inc.* should be overturned[93] and platforms be subject to common-law distributor liability.[94] This would in effect create a liability on notice regime, much like the notice-and-takedown system of § 512. Similarly, the Platform Accountability and Consumer Transparency (PACT) Act would have created a system for material that a court had determined to be unlawful and would have defined a platform to have knowledge only when it was provided with a copy of the court order and information reasonably sufficient to locate the material.[95] This is a substantial improvement on the deficiencies of notice-and-takedown under § 512 because it sets a meaningful threshold for sending an effective notice. On the other hand, the process of obtaining a court order will be slow and expensive, so this would be a solution only for egregiously harmful material.

B. SECTION 512

Our model sheds light on the notice-and-takedown regime of § 512 of the Copyright Act.[96] The basic rule of § 512 is that a hosting platform "shall not be liable for monetary relief . . . for infringement of copyright by reason of the storage at the direction of a user of [infringing] material."[97] This is a blanket immunity, but it is qualified by five (!) exceptions.

First, § 512(c)(1)(A) removes the platform's immunity as to specific material if it has "actual knowledge that the material . . . is infringing"[98] and the platform does not "act[] expeditiously to remove, or disable access to, the material."[99] This exception reflects the intuition that where the platform *has* performed an investigation into specific content, it can remove harmful items without affecting non-harmful items. It does not matter where along the $\lambda(x)$ curve the item falls; once the platform has knowledge, it must act.

93. Zeran v. Am. Online, Inc., 129 F.3d 327 (4th Cir. 1997) (holding that Section 230 provides a blanket immunity from defamation liability for platforms carrying third-party content).

94. *See, e.g.*, Shlomo Klapper, *Reading Section 230*, 70 BUFF. L. REV. 1237 (2022).

95. *See* S. 4066, 116th Cong. (2020); *see also* Daphne Keller, *CDA 230 Reform Grows Up: The PACT Act Has Problems, but It's Talking About the Right Things*, STAN. L. SCH.: CTR. FOR INTERNET & SOC'Y (July 16, 2020), https://cyberlaw.stanford.edu/blog/2020/07/cda-230-reform-grows-pact-act-has-problems-it%E2%80%99s-talking-about-right-things.

96. *See* 17 U.S.C. § 512.

97. 17 U.S.C. § 512(c)(1).

98. *Id.* § 512(c)(1)(A)(i).

99. *Id.* § 512(c)(1)(A)(iii).

Second, as discussed above, § 512(c)(1)(A)(ii) removes the platform's immunity as to specific content if it is "aware of facts or circumstances from which infringing activity is apparent" and fails to remove the content.[100] This is a negligence rule.

Third, § 512(c)(1)(B) removes the platform's immunity if it "receive[s] a financial benefit directly attributable to the infringing activity, in a case in which the service provider has the right and ability to control such activity."[101] This standard, which resembles but is not identical in application to the common-law vicarious-infringement standard,[102] is not in theory tied to the platform's knowledge at all. Instead, it is designed to smoke out situations in which a platform that could block infringement has especially bad incentives to turn a blind eye to it. In terms of our model, we think these are situations in which c is small (so that the platform has the "ability to control" infringement) and P is large (so that the platform has strong private incentives to allow as much infringement as it can). These are circumstances under which in the absence of liability, the platform might under-invest in investigating likely-to-be-infringing content.

Fourth, § 512(c)(1)(C) removes the platform's immunity if it receives a "notification of claimed infringement" and fails to remove it.[103] As discussed above, this creates a notice-and-takedown regime, which is effective only to the extent that sending a notice is a signal that conveys information.

And fifth, again as discussed above, the repeat-infringer provision of § 512(i) creates a conditional immunity.[104]

To summarize, the five limitations on the § 512(c) safe harbor all function in different ways. The actual-knowledge provision deals with cases where $c = 0$ and no investigation is required; the red-flag provision deals with cases where $\lambda(x)$ is high and the content is likely to infringe; the financial-benefit provision deals with cases where c is low and P is high so the platform has bad incentives not to investigate; the notice-and-takedown provision deals with cases where the copyright owner has taken on the investigative costs c itself; and the RIP provision requires the platform to keep overall infringement beneath a total threshold. Notably, four out of these five limitations have to do with investigation costs.

100. *Id.* § 512(c)(1)(A)(ii).

101. *Id.* § 512(c)(1)(B).

102. *See* R. Anthony Reese, *The Relationship Between the ISP Safe Harbors and the Ordinary Rules of Copyright Liability*, 32 COLUM. J.L. & ARTS 427 (2009).

103. 17 U.S.C. § 512(c)(1)(C).

104. *See supra* Section IV.D.

Section 512 also has important text on investigation. The safe harbor does not depend on "a service provider monitoring its service or affirmatively seeking facts indicating infringing activity."[105] A useful way to understand this statement is as creating a rule that a platform's choice of whether to investigate content is not a basis for liability. Only the fact that the platform hosts infringing content is a liability trigger, and the safe harbor is removed only when the platform's conduct falls into one of the five limitations above. These are all performance standards based on the platform's knowledge or activity *with respect to the infringing content*. The platform is free to arrange its activities as it chooses, investigating only the content it chooses to, as long as it acts when it has knowledge.

C. THE DIGITAL SERVICES ACT

The Digital Services Act (DSA) makes a number of interesting choices.[106] The first is that it sharply distinguishes between platforms that serve as a "mere conduit" and those that store information at the request of a user. A mere conduit is not liable for user-provided content and has no content-moderation obligations.[107] But a hosting service is liable only when it has knowledge (actual or red-flag) and fails to act.[108] Thus, mere conduits have a blanket immunity, while hosting services have a notice-and-takedown regime.[109]

Above, we criticized the two-track regime under pre-§ 230 law for creating a disincentive for platforms to engage in content moderation. The DSA's distinction is more sensible because it is tied to the nature of a platform's services rather than the nature of its moderation. A platform can qualify for the mere-conduit safe harbor (or the similar safe harbor for caching services[110]) only when it is completely passive with respect to the information, selecting neither the material nor its destination and playing no role in modifying the material.

The DSA's hosting safe harbor is in some respects broader than the safe harbor under § 512. It has actual-knowledge and red-flag exceptions, but it does not have anything that looks like the vicarious-liability provision of § 512

105. 17 U.S.C. § 512(m)(1).

106. *See* Regulation (EU) 2022/2065, of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and Amending Directive 2000/31/EC (Digital Services Act), 2022 O.J. (L. 277).

107. *See id.* art. 4.

108. *See id.* art. 6.

109. Section 512 draws a similar distinction, but only applies to copyright infringement, whereas the DSA applies to all "illegal activity or illegal content." *See id.* art. 6(1).

110. *See id.* art. 5.

or the conditional immunity of the RIP. While there is a requirement that platforms suspend service to users "that frequently provide manifestly illegal content," this is simply an independent requirement of law, not a condition on the safe harbor.[111] It therefore avoids some of the error costs and overdeterrence associated with the RIP. The DSA has a takedown procedure that is based not on private notices but on orders from appropriate authorities, which functions as a high-threshold notice-and-takedown procedure.[112]

The DSA also has a separate procedure for "notice and action mechanisms" that allow private parties to send notices that

> shall be considered to give rise to actual knowledge or awareness for the purposes of Article 6 in respect of the specific item of information concerned where they allow a diligent provider of hosting services to identify the illegality of the relevant activity or information without a detailed legal examination.[113]

This is a U.S.-style notice-and-takedown regime. And, most interestingly, it has a "trusted flagger" provision that allows member states to designate certain entities as trusted flaggers, whose notices of illegal material "are given priority and are processed and decided upon without undue delay."[114] A trusted flagger is required to act "diligently, accurately and objectively" and should therefore not send notices without suitable investigation.[115] This is a clever response to the signaling problem we discuss above with respect to notice-and-takedown under § 512.

The DSA emphasizes that platforms remain eligible for the safe harbors even if they "in good faith and in a diligent manner, carry out voluntary own-initiative investigations into, or take other measures aimed at detecting, identifying and removing, or disabling access to, illegal content."[116] This is an important limitation to prevent the *Stratton Oakmont* trap discussed in Section II.G above. Contrariwise, it adds that platforms have "[n]o general obligation to monitor the information" they carry "nor actively to seek facts or circumstances indicating illegal activity."[117] This is (like the corresponding provision in § 512) a way of emphasizing that the platform can choose not to investigate content and that those choices by themselves do not create liability.

111. *Id.* art. 23(1).
112. *See id.* art. 9.
113. *Id.* art. 16.
114. *Id.* art. 22.
115. *Id.*
116. *Id.* art. 7.
117. *Id.* art. 8.

VII. CONCLUSION AND FUTURE EXTENSIONS

Our model is deliberately simple. Nonetheless, it yields vivid, straightforward intuitions about a wide range of intermediary-liability problems. We hope that it can provide a clean foundation for modular extensions to model a wider range of fact patterns and legal responses.

Indeed, our treatment of liability on notice is intended as an example of how to extend our basic model. We introduced a parsimonious extension: a new type of actor (victims of harm), who can take two types of actions (investigate and give notice), and whose features are captured by a single parameter (their cost of investigation c_v). A more sophisticated treatment of liability on notice might add costs of sending notices (or of sending false notices) and allow victims and platforms to negotiate deals.

Other extensions of our model might introduce other actors, such as the users who post content in the first place. One could posit, for example, that these users know whether the content they are posting is harmful or not and have private gains from posting that are distinct from the platform's revenues but do not exhaust the social benefits their posting creates. Add in a feature to model the comparative difficulty of seeking enforcement against these users, and again, one is in a position to draw interesting conclusions. Perhaps these users might negotiate the price they pay for posting on the platform, or perhaps the platform competes with other platforms, and so on.

Another way in which the model presented in this Article might be limited is the assumption that harmful content is less profitable and has fewer positive spillovers. We made this assumption because it simplifies the analysis in our initial presentation. Our results would be robust if, for example, $p(x)$ is increasing but $\lambda(x)h$ increases faster than $p(x)$ (as measured by the slope). In other cases, however, it becomes possible for $s(x)$ and $\lambda(x)h$ to intersect multiple times—even infinitely often—and it is no longer rational for a moderator to set a single threshold x_s. Instead, as $s(x)$ and $\lambda(x)h$ take turns surging ahead, the moderator might choose to turn moderation on and off repeatedly. A similar point applies to the platform's revenues $p(x)$, and one might also consider whether the harms h and costs of investigation c should vary.

Although the analysis will be more mathematically difficult, there are important classes of content for which $s(x)$ and $p(x)$ plausibly increase even as the content becomes more likely to be harmful. The most scandalous accusations against public figures are both more likely to be false and more important to air publicly if true. Indeed, our analysis of § 512 vicarious liability suggests that it makes the most sense in a world where $p(x)$ increases with

$\lambda(x)$. The fact that space limitations prevent us from addressing this scenario in the depth a proper analysis would require should not be taken as a statement that the scenario does not occur or is not worth understanding when it does.

Another important set of extensions relates to error costs. We have considered errors by *platforms* about whether content is harmful. But our model assumes that courts eventually reach the truth. This assumption may not be warranted because courts themselves have an error rate and may classify harmful content as harmless or vice versa. In our discussion of negligence and conditional immunity, we noted that courts and regulators may mismeasure factors that go into setting and applying liability thresholds. If a court misunderstands the threshold or a platform's efforts, the consequences can be significant. Platforms must make their moderation decisions in the shadow of the possibility that courts could err.