Michigan Technology Law Review

Article 2

2021

Content Moderation Remedies

Eric Goldman
Santa Clara University School of Law, egoldman@gmail.com

CONTENT MODERATION REMEDIES

*Eric Goldman**

ABSTRACT

This Article addresses a critical but underexplored aspect of content moderation: if a user's online content or actions violate an Internet service's rules, what should happen next? The longstanding expectation is that Internet services should remove violative content or accounts from their services as quickly as possible, and many laws mandate that result. However, Internet services have a wide range of other options—what I call "remedies"—they can use to redress content or accounts that violate the applicable rules. This Article describes dozens of remedies that Internet services have actually imposed. It then provides a normative framework to help Internet services and regulators navigate these remedial options to address the many difficult tradeoffs involved in content moderation. By moving

* Associate Dean for Research, Professor of Law, and Co-Director of the High Tech Law Institute, Santa Clara University School of Law. Email: egoldman@gmail.com. Website: http://www.ericgoldman.org.

I am a founding board member of the Trust & Safety Professional Association (for content moderation professionals) and the Trust & Safety Foundation. I was General Counsel of Epinions.com, a consumer review service, from 2000-02.

I appreciate the helpful comments from participants at USENIX Free and Open Communications on the Internet Workshop (FOCI '19); the Seventh Annual Computer Science and the Law Workshop at the University of Pennsylvania Law School; the Princeton University Center for Information Technology Policy Luncheon Series; the 9th Annual Internet Law Works-in-Progress Conference; the Works in Progress in Intellectual Property (WIPIP) at University of Houston Law Center; the Faculty Workshop at Santa Clara University School of Law; the Tel Aviv University Faculty of Law Technology Law Seminar; the Law and Technology Scholarship Seminar at UC Berkeley School of Law; the Faculty Workshop at University of San Diego School of Law; Center for Intellectual Property Research's International IP Colloquium at Indiana University Maurer School of Law; the Freedom of Expression Scholars Conference 2020 at Yale Law School; the University of Colorado School of Law Content and Platforms Seminar; and the Regulation at Scale Virtual Roundtable at University of Nebraska School of Law; as well as Colleen Chien, Julie Cohen, Thomas DeGuzman, Evelyn Douek, Eli Edwards, Alex Feerst, James Grimmelmann, Tomiwa Ilori, Thomas Kadri, Daphne Keller, Aleksandra Kuczerawy, Orly Lobel, Alex Macgillivray, Judy Malloy, Enguerrand Marique, Yseult Marique, Jess Miers, Irina Raicu, Lisa Ramsey, Betsy Rosenblatt, Colin Rule, Jessica Silbey, David Sloss, Rebecca Tushnet, Rachel Wolbers, Tseming Yang, and Tal Zarsky.

This project was supported in part by grants from the John S. and James L. Knight Foundation and the Nebraska Governance and Technology Center's Summer Grant Program 2020.

past the binary remove-or-not remedy framework that dominates the current discourse about content moderation, this Article helps to improve the efficacy of content moderation, promote free expression, promote competition among Internet services, and improve Internet services' community-building functions.

TABLE OF CONTENTS

INTRODUCTION

In May 2019, a supporter of President Trump published a manipulated video of House Speaker Nancy Pelosi that slowed down authentic footage

while maintaining the original voice pitch,[1] creating the false impression that Speaker Pelosi had delivered her remarks while intoxicated. The video became a viral sensation and spread rapidly across the Internet.[2]

The hoax video raises many interesting policy questions,[3] including how the three major social media services (Facebook, Twitter, and YouTube) responded to the video. The video, though misleading, probably did not constitute defamation or otherwise violate the law;[4] and even if it did, the social media services likely did not face any legal exposure from it.[5] As a result, the social media services had the legal freedom to moderate the video as they saw fit.

1. Kevin Poulsen, *We Found the Guy Behind the Viral 'Drunk Pelosi' Video*, DAILY BEAST (June 2, 2019, 11:14 PM), https://www.thedailybeast.com/we-found-shawn-brooks-the-guy-behind-the-viral-drunk-pelosi-video.

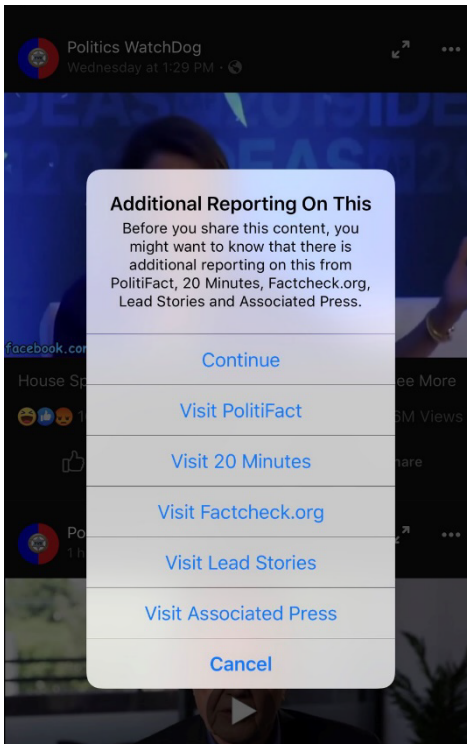
2. Sue Halpern, *Facebook's False Standards for Not Removing a Fake Nancy Pelosi Video*, NEW YORKER (May 28, 2019), https://www.newyorker.com/tech/annals-of-technology/facebooks-false-standards-for-not-removing-a-fake-nancy-pelosi-video. One version of the video was viewed two million times.

3. For example, there are substantial and legitimate concerns about authentic-looking "deepfake" videos. *See* Robert Chesney & Danielle Keats Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, 107 CALIF. L. REV. 1753, 1784–85 (2019). The Pelosi video was a "cheap fake," a euphemism for manipulated authentic videos. BRITT PARIS & JOAN DONOVAN, DATA & SOC'Y, DEEPFAKES AND CHEAP FAKES: THE MANIPULATION OF AUDIO AND VISUAL EVIDENCE 24 (2019), https://datasociety.net/wp-content/uploads/2019/09/DS_Deepfakes_Cheap_FakesFinal-1.pdf.

4. The video was likely constitutionally protected as political commentary.

5. 47 U.S.C. § 230 says that websites are not liable for third-party content such as the hoax video. *See* Eric Goldman, *An Overview of the United States' Section 230 Internet Immunity*, *in* THE OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY 155 (Giancarlo Frosio ed., 2020) [hereinafter Goldman, *Section 230 Overview*].

Often, these three services reach the same conclusions about how to handle controversial high-profile content. But, not in this case. Instead, each service did something different with the Pelosi hoax video. Twitter left the video up.[6] YouTube removed the video.[7] Facebook allowed the video to remain on its service but attempted to dissuade users from sharing it[8] by adding the disclaimer seen below.[9]



Facebook received heavy criticism for not removing the video,[10] but its decision raises intriguing possibilities. Ordinarily, we assume that social media and other services publishing third-party content make a binary

6. Halpern, *supra* note 2.

7. Emily Stewart, A *Fake Viral Video Makes Nancy Pelosi Look Drunk. Facebook Won't Take It Down*., VOX: RECODE (May 24, 2019, 3:50 PM), https://www.vox.com/recode/2019/5/24/18638822/nancy-pelosi-doctored-video-drunk-facebook-trump.

8. Halpern, *supra* note 2.

9. Donie O'Sullivan (@donie), TWITTER (May 25, 2019, 12:47 PM), https://twitter.com/donie/status/1132327255802294274. If you cannot read the photo, Facebook's pop-up warning says: "Before you share this content, you might want to know that there is additional reporting on this from PolitiFact, 20 Minutes, Factcheck.org, Lead Stories and Associated Press" with links to each of those sources.

10. *E.g.*, Donie O'Sullivan, *Pelosi Calls Facebook a 'Shameful' Company That Helped in 'Misleading the American People'*, CNN BUSINESS (Jan. 16, 2020, 1:21 PM), https://www.cnn.com/2020/01/16/tech/pelosi-shameful-facebook/index.html.

choice: leave content up (like Twitter did) or remove it (like YouTube did). Facebook chose a different option. That prompts the questions: what other alternative options are available, and when might they be better than the standard binary options?

* * *

How Internet services that publish third-party content ("Internet services")[11] decide to publish or remove third-party content—a process called content moderation[12]—has become a major issue in our society, and for good reason. As the Pelosi hoax video example shows, an Internet service's decision can have major political implications. Other content moderation decisions can have dramatic—even life-changing—consequences for authors, victims, and many others.

Due to the high stakes, the conventional wisdom is that when online user content or accounts violate the applicable rules,[13] they should be removed as quickly as possible,[14] especially if the service has been notified of the problem. I refer to this as the "removal" remedy or the "binary" approach to redressing violations (i.e., a content moderation decision functions like an on/off switch). Many laws around the world have codified the binary approach to remedies.[15]

Unfortunately, the presumption of "removal"[16] has overshadowed other ways to redress violative online content and activity. Nevertheless, facilitated by the legal freedom provided by Section 230's immunity for

---

11. This Article applies to all Internet services that gather, organize, and publish third-party content, including user-generated content ("UGC") services and platforms.

12. *See, e.g.*, *What is Content Moderation?*, BESEDO (Nov. 20, 2020), https://besedo.com/resources/blog/what-is-content-moderation ("Content moderation is when an online platform screen and monitor user-generated content based on platform-specific rules and guidelines to determine if the content should be published on the online platform, or not."); James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 47 (2015) (defining moderation as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse"); Shagun Jhaver, Amy Bruckman & Eric Gilbert, *Does Transparency in Moderation Really Matter? User Behavior After Content Removal Explanations on Reddit*, 3 PROC. OF THE ACM ON HUM.-COMPUT. INTERACTION, Nov. 2019, at 1, 4 ("Content moderation determines which posts are allowed to stay online and which are removed, how prominently the allowed posts are displayed, and which actions accompany content removals.").

13. As discussed in Part I(A), this Article generally does not distinguish between "illegal" content/activity and content/activity that violates an Internet service's "house rules." Alternative remedies could help in both circumstances. However, Part IV(A)(1) will address how the severity of a rule violation might influence the remedial determinations, and illegality often will be more severe than house rule violations.

14. Removals can be global across a service's entire network or done for only specific geographies or portions of the network. Parts II(A)(3) and IV(B)(5) revisit these differences.

15. *See infra* Part II for examples of such laws.

16. *See* MacKenzie F. Common, *Fear the Reaper: How Content Moderation Rules Are Enforced on Social Media*, 34 INT'L REV. L. COMPUT. & TECH. 126, 129–31 (2020).

content moderation decisions,[17] Internet services have experimented with and deployed many alternative remediation techniques in the past few years.[18]

This Article addresses this underexamined phenomenon through two successive inquiries. First, the Article comprehensively describes and organizes dozens of "remedies" that Internet services have used to redress user violations. Then, the Article turns to the normative questions: how should these remedy options be prioritized, which remedies are best, and why?

This Article advances the discourse about content moderation in two important ways. First, the Article documents the range of diverse remedies that are potentially available. As Internet services experiment with different options, they can find new and better ways to balance the often-difficult policy tradeoffs inherent in content moderation, such as how to remediate anti-social online content or behavior while still advancing free expression. Second, the Article shows how Internet services can adopt idiosyncratic remedial strategies, increasing the potential bases of competitive differentiation and allowing them to serve their unique audiences better.

Internet services can only achieve the full potential of alternative remedial options if regulators let them. This may be unrealistic. To date, when regulators have specified remedies for legal violations, they routinely have mandated removal as the sole remedy for user violations, thereby eliminating Internet services from using their discretion to explore the full spectrum of potential remedies.

The process of content moderation has significant implications for how we engage and communicate with each other as a society. Limiting the range of remedies available to redress violative content hinders our ability to optimize and fine-tune content moderation processes and achieve these socially important goals.

The Article proceeds in four parts. Part I explains how this Article fits into the content moderation and remedies literatures. Part II demonstrates how the leave up/remove binary remedial approach is hard-wired into the law and discourse and why we would benefit from moving past it. Part III provides a comprehensive inventory of content moderation remedies. Part IV explores how Internet services and regulators can navigate the options enumerated in Part III to advance various normative goals. A short

17. 47 U.S.C. § 230.

18. In a recent example, Internet services reacted to the Capitol insurrection of January 6, 2021 with a wide range of remedies, including the typical content and account removals and suspensions as well as specialized remedies such as banning certain phrases in hashtags and eliminating a Twitch emote. *See Platform Actions in Response to January 6 Capitol Events—Newest to Oldest*, FIRST DRAFT, https://docs.google.com/document/d/1dNC87RtdPWBXXReTsrAl-Sknw4PtwanPX0CA_oi20ec/edit?fbclid=IwAR05m4XHSS-H2znFuSKIlGhBN0FnQqUoqgxau0vbQOST-yEMF9aCnnoPM9w# (Jan. 16, 2021).

conclusion addresses why regulators almost certainly will force Internet services down the worst path.

I. PROJECT CONTEXT

This Part explains this Article's relationship with the existing content moderation and remedies literatures.

A. *Relationship to the Content Moderation Literature*

The social importance of content moderation has spurred a robust academic conversation,[19] supplemented by an even more active academic conversation about related topics such as "platform governance" and "algorithmic accountability." Collectively, this literature generally addresses one of three topics:

Topic 1: What content and activity should be allowed online? These are the substantive rules for content and activities, such as rules that child pornography and copyright infringement are not permissible or that political speech is generally permitted. There are longstanding, ongoing, and vigorous debates over what content and activities should be permitted online.

Topic 2: Who should make the substantive rules of online content and activities? Rulemaking is a core function of government, which expresses its rules through official substantive law—such as legislatures or courts determining that certain content and activities are illegal or tortious—or "soft" law, such as when regulators cajole Internet companies[20] to "voluntarily" redress "lawful but awful" content.[21]

19. *E.g.*, TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 176 (2018); SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019); NICOLAS P. SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES (2019); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

20. Derek E. Bambauer, *Against Jawboning*, 100 MINN. L. REV. 51, 74–78 (2015). This can also be called "working the ref." *E.g.*, Eric Alterman, *The Right Is Working the Ref Yet Again. This Time on Facebook—and It's Working*, NATION (Aug. 16, 2018), https://www.thenation.com/article/archive/the-right-is-working-the-ref.

21. Eric Goldman & Jess Miers, *Online Account Terminations/Content Removals and the Benefits of Internet Services Enforcing Their House Rules*, 1 J. FREE SPEECH L. 191, 194 (2021). For example, the U.K. wants Internet services to combat harmful content, even if it is lawful. *See* JEREMY WRIGHT & SAJID JAVID, ONLINE HARMS WHITE PAPER (2019); HOME DEPARTMENT & DEPARTMENT FOR DIGITAL, CULTURE, MEDIA & SPORT, ONLINE HARMS WHITE PAPER: FULL GOVERNMENT RESPONSE TO THE CONSULTATION, 2020, Cm. 354 (UK); Eric Goldman, *The U.K. Online Harms White Paper and the Internet's Cable-ized Future*, 16 OHIO ST. TECH. L.J. 351 (2020) [hereinafter Goldman, *UK Online Harms*].

Companies may voluntarily adopt their own substantive rules for content and activities on their services, what I call "house rules."[22] House rules supplement the government-created rules by restricting otherwise-legal content or activities based on their idiosyncratic editorial policies.[23]

Topic 3: Who should determine if a rule violation has occurred, and who should hear any appeals of those decisions? Historically, courts or other government entities have played a preeminent role in adjudicating rule violations, at least with respect to matters important enough to justify the high adjudication costs. In contrast, with respect to online content or actions, Internet services make their own determinations of whether a rule violation has occurred, although sometimes they may choose to honor the decisions of independent third parties.[24]

This Article does not directly address any of the prior three topics. Instead, this Article focuses on a fourth topic that has received comparatively less attention: after a rule violation has occurred,[25] what steps ("remedies") should the service take to redress the violation?[26]

Admittedly, it is hard to discuss remedies for rule violations independently of the other three topics. The legitimacy of any remedy will depend, in part, on the legitimacy of the underlying content moderation system, including the rules, who set them, and how violations were

---

22. Google/YouTube calls them "rules of the road," including their "content policies" and "community guidelines." GOOGLE, INFORMATION QUALITY & CONTENT MODERATION 6 (2020), https://blog.google/documents/83/information_quality_content_moderation_white_paper.pdf [hereinafter YouTube Report].

23. *See* Goldman & Miers, *supra* note 21, at 194–95.

24. Two examples:

- Ripoff Report's Arbitration Program lets third-party arbitrators redact portions of negative Ripoff Report reviews. *VIP Arbitration Program*, RIPOFF REPORT, https://www.ripoffreport.com/arbitration (July 1, 2020).
- Facebook honors the decisions of its Oversight Board (sometimes called the "Facebook Supreme Court"). OVERSIGHT BD., https://oversightboard.com/ (last visited Oct. 21, 2021); Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418 (2020).

"Social Media Councils" would act like Facebook's Oversight Board. *See*, *e.g.*, ARTICLE 19, THE SOCIAL MEDIA COUNCILS: CONSULTATION PAPER (2019), https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf; TRANSATLANTIC WORKING GRP., FREEDOM AND ACCOUNTABILITY: A TRANSATLANTIC FRAMEWORK FOR MODERATING SPEECH ONLINE 26–27 (2020), https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf.

25. Until Part IV, this Article treats all crimes, torts, and violations of house rules as equally appropriate triggers for ex post remedies.

26. *See* DOUGLAS LAYCOCK & RICHARD L. HASEN, MODERN AMERICAN REMEDIES: CASES & MATERIALS 2 (Concise 5th ed. 2018) ("In every case, we will assume that defendant's conduct is unlawful and ask what the court can do about it: What does plaintiff get? How much does he get? Why does he get that instead of something more, or less, or entirely different?").

determined. If the content moderation scheme lacks legitimacy, any associated remedies will too. The close interplay between the content moderation process and the associated remedies means that isolating the remedies can feel incomplete. Inevitably, a discussion of the remedies migrates back to aspects of the other three topics.

Nevertheless, isolating the remedies topic helps spotlight an issue that otherwise gets overshadowed. It also means this Article can analyze the remedial issues more thoroughly than if it tried to comprehensively engage the full range of content moderation topics.[27] Part IV will relax this constraint and reconsider other aspects of content moderation.

B. *Relationship to the Remedies Literature*

This Article focuses on what happens after an item of online content or an online account has been determined to violate the applicable rules. This Article calls those ex post consequences "remedies," because the responses are intended to remediate the rule violation, in the same way that a court grants remedies to successful litigants who are entitled to legal relief. These ex post responses could be called "sanctions,"[28] "penalties," or "punishments," but this would exclude the many non-punitive remedies discussed in Part III.[29]

There is a rich and venerable academic literature about remedies for legal violations. For example, the criminal justice system encodes policy goals such as punishment/retribution, deterrence, incapacitation (segregating dangerous individuals from the community), rehabilitation, expressive justice, and victim restitution.[30] Those normative values should influence content moderation design as well.[31]

---

27. *See id.* at 7 ("Whether we design remedies that encourage profitable violations, or remedies that seek to minimize violations, or remedies that serve some other purpose altogether, we are making choices distinct from the choices we make when we design the rest of the substantive law.").

28. Marique & Marique define "sanctions" as "the exercise of power and taken by digital operators towards undesirable behavior on the modern public square. Sanctions react to a specific problematic behavior defined as such by a socially recognized rule." Enguerrand Marique & Yseult Marique, *Sanctions on Digital Platforms: Balancing Proportionality in a Modern Public Square*, 36 COMPUT. L. & SEC. REV., Apr. 2020, at 1, 5 (2020). "Sanctions" is also the nomenclature used in the international treaty compliance literature. *See, e.g.*, ABRAM CHAYES & ANTONIA HANDLER CHAYES, THE NEW SOVEREIGNTY: COMPLIANCE WITH INTERNATIONAL REGULATORY AGREEMENTS (1995).

29. Internet services sometimes use "action" as a verb for their content decisions. For example, Pinterest explained that a removed group "was actioned and labeled for misinformation, specifically conspiracies and health misinformation." Jason Koebler, *Pinterest Bans Anti-Abortion Group Live Action for Posting Misinformation*, VICE (June 12, 2019, 10:42 AM), https://www.vice.com/en_us/article/ywyx7g/pinterest-bans-anti-abortion-group-live-action-for-posting-misinformation.

30. *E.g.*, SANFORD H. KADISH & STEPHEN J. SCHULHOFER, CRIMINAL LAW AND ITS PROCESSES: CASES & MATERIALS 101–53 (6th ed. 1995); JOSHUA DRESSLER, UNDERSTANDING

Nevertheless, this Article addresses fundamentally different issues than the standard remedies literature. At its core, this Article focuses on editorial decisions implemented by private entities, not decisions made by government state actors.[32] This difference matters:

Accountability. The government imposes its rules on its citizens, whether they agree or not, though it must give citizens fair notice of the rules. Citizens must honor the government-set rules that apply to them, and usually they pay taxes to fund government services such as a judicial system. Citizens get a voice in this governance through their right to vote.

Private companies are categorically different. They cannot impose taxes, compel rule compliance through tax-funded police powers, or be voted out in elections. Most importantly, they cannot compel citizens to use them. As a result, the remedy schemes of Internet services have different accountability mechanisms and different impacts than those imposed by governments.

Unavailability of Certain Remedies. Many remedies available to state actors are categorically unavailable to Internet services.[33] For example,

---

CRIMINAL LAW, 11–26 (7th ed. 2015); *see also* Randy E. Barnett, *Restitution: A New Paradigm for Criminal Justice*, 87 ETHICS 279 (1977); Richard A. Posner, *Retribution and Related Concepts of Punishment*, 9 J. LEG. STUD. 71, 71 (1980). Congress codified some of these normative values into federal criminal sentencing:

> The court, in determining the particular sentence to be imposed, shall consider—
> (2) the need for the sentence imposed—
>
> > (A) to reflect the seriousness of the offense, to promote respect for the law, and to provide just punishment for the offense;
> > (B) to afford adequate deterrence to criminal conduct;
> > (C) to protect the public from further crimes of the defendant; and
> > (D) to provide the defendant with needed educational or vocational training, medical care, or other correctional treatment in the most effective manner;
> > . . .
>
> (7) the need to provide restitution to any victims of the offense.

18 U.S.C. § 3553(a).

Civil remedial goals are similar. Laycock & Hasen categorize civil remedies as compensatory remedies, preventive remedies (including coercive and declaratory remedies), restitutionary remedies, punitive remedies, and ancillary remedies. LAYCOCK & HASEN, *supra* note 26, at 2. *See generally* Marco Jimenez, *Remedial Consilience*, 62 EMORY L.J. 1309 (2013) (describing remedial interests of restoration, retribution, coercion, and protection); Marique & Marique, *supra* note 28, at 7–8 (discussing how "sanctions" can be "retributive," "reparative," or "pedagogic").

31. *E.g.*, Sarita Schoenebeck, Oliver L Haimson, & Lisa Nakamura, *Drawing from Justice Theories to Support Targets of Online Harassment*, 23 NEW MEDIA & SOC'Y 1278 (2020) (discussing how criminal justice theories can inform content moderation).

32. *E.g.*, LAYCOCK & HASEN, *supra* note 26, at 1 ("A remedy is anything *a court* can do for a litigant who has been wronged or is about to be wronged.") (emphasis added).

33. Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 806–07 (2021) ("[W]hile there can be serious consequences from having content or accounts removed from social media, these will usually fall short of the consequences of state sanction.").

Internet services cannot directly garnish a person's wages;[34] seize their physical assets; remove a child from a parent's custody; shoot tear-gas at peaceful protestors; incarcerate a person or otherwise deprive them of their physical freedom; or impose capital punishment.

Internet services also can only regulate behavior within their virtual "premises."[35] Because the intersection between the service's virtual premises and a non-compliant user's activities or assets may be relatively limited, an Internet service has a far more limited toolkit of remedy options than government actors who can reach virtually every aspect of a person's life.

The Laws of Nature Do Not Apply. Governments' coercive powers are intrinsically constrained by the laws of nature. For example, governments cannot incarcerate a person who is not physically present. In contrast, physics do not apply to Internet services' remedies; those remedies are constrained solely by the technical limits of the underlying software code.[36] For example, Internet services can turn a game player's avatar into a virtual toad with restricted functionality (called "toading").[37] Due to the laws of nature, there is no offline equivalent remedy to toading. Freed from the laws of nature, Internet services can create and implement remedies that have no offline analogues.

Constitutional Limits. Because governments have extraordinary police powers that citizens cannot reject, the Constitution protects citizens from abuses of the government's coercive powers.[38] Due to their fundamentally different role in our society, private entities are not subject to these Constitutional restrictions. Indeed, courts routinely reject efforts to impose Constitutional obligations on Internet companies predicated on the argument that they are like the government.[39]

---

34. However, services that compensate their users can stop paying, an option considered in Part III.

35. *See, e.g.*, Jennifer L. Mnookin, *Virtual(ly) Law: The Emergence of Law in LambdaMoo: Mnookin*, 2 J. COMPUT.-MEDIATED COMMC'N (1996) (discussing how LambdaMOO intentionally limited its remedial system to in-world consequences).

36. LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

37. *See generally* Mnookin, *supra* note 35, at n.44.

38. *See Developments in the Law: Alternatives to Incarceration*, 111 HARV. L. REV. 1863, 1950–55 (1998) [hereinafter *Incarceration Alternatives*] (discussing constitutional challenges to incarceration alternatives).

39. "[C]ase law has rejected the notion that private companies such as Facebook are public fora . . . . [S]imply because Facebook has many users that create or share content, it does not mean that Facebook . . . becomes a public forum." Federal Agency of News LLC v. Facebook, Inc., 432 F. Supp. 3d 1107, 1121–22 (N.D. Cal. 2020); *see also* Prager Univ. v. Google LLC, 951 F.3d 991, 995–99 (9th Cir. 2020); Divino Grp. LLC v. Google LLC, No. 19-cv-04749-VKD, 2021 WL 51715, at *4–7 (N.D. Cal. Jan. 6, 2021); Buza v. Yahoo!, Inc., No. C 11–4422 RS, 2011 WL 5041174 at *1 (N.D. Cal. Oct. 24, 2011); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 631–32 (D. Del. 2007); Eric Goldman, *Of Course the First Amendment Protects Google and Facebook (and It's Not a Close Question)* (Santa Clara

The "remedies" academic literature generally assumes that state actors will determine and implement the remedies. Private actors, with their different structural attributes, raise different considerations that do not fit with the standard remedies literature.[40] The divergent structural attributes of government and private actors necessitates different analytical tools.

II. THE UBIQUITY OF THE REMOVALS REMEDY

This Part demonstrates the pervasiveness of the binary approach to online remedies and then discusses the benefits of thinking more broadly.

A. *The Historical Embrace of the Binary Approach to Remedies*

Regulators have codified removals as the primary or exclusive remedy in many laws throughout the world.[41] Similarly, civil society entities have issued principles to help guide the development of Internet law, and those principles also reflect binary thinking about remedies. This subpart documents seven examples of the pervasiveness of the binary approach to content moderation remedies:

1. DMCA Online Safe Harbors

In 1998, Congress sought to update copyright law for the digital age, and the era of user-generated content, in a law called The Digital Millennium Copyright Act (DMCA).[42] The DMCA included a safe harbor for hosting user-generated content, codified at § 512(c) of the Copyright Act.[43] This safe harbor incorporated the binary remedies of both content removal and account termination[44]:

Content Removal. The § 512(c) safe harbor contemplates that copyright owners will notify services of allegedly infringing user uploads.[45] To obtain the safe harbor, the services then must expeditiously "remove[] or disable access to"[46] user-uploaded files in response to the copyright owners'

Legal Rsch. Paper, Paper No. 08, 2018), https://ssrn.com/abstract=3133496; Ashutosh Bhagwat, *Do Platforms Have Editorial Rights?*, 1 J. FREE SPEECH L. 97 (2021).

40. *See* Maayan Perel, *Digital Remedies*, 35 BERKELEY TECH. L.J. 1, 37–42 (2020) (discussing the problems when courts delegate responsibility for implementing equitable relief to private Internet companies).

41. *Cf.*, Dan M. Kahan, *What Do Alternative Sanctions Mean?*, 63 U. CHI. L. REV. 591 (1996); *Incarceration Alternatives*, *supra* note 38 (providing an analogous discussion to the binary approach to remedies encoded into criminal remedies).

42. Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (1998) (codified in scattered sections of 17 U.S.C. & 28 U.S.C.).

43. 17 U.S.C. § 512(c).

44. Grimmelmann, *supra* note 12, at 107.

45. 17 U.S.C. § 512(c)(3).

46. *Id.* § 512(c)(1)(C). The statute does not explain the differences between removal and disabling access. Ian Ballon says that "[t]here are legitimate reasons why a service

notice.[47] This provision is commonly called the "notice-and-takedown" provision.

Account Termination. To be eligible for the § 512 safe harbors, Internet services must reasonably implement policies to terminate "repeat infringers."[48] To identify recidivists, services must track infringing users[49] and issue "strikes."[50] The safe harbor also requires services to terminate user accounts that receive too many strikes[51] (though the statute does not specify the exact number of strikes that cause a user to be a "repeat" infringer).[52]

2. E.U. E-Commerce Directive and Its Progeny

Soon after Congress enacted the DMCA, the European Union adopted its "E-Commerce Directive."[53] Like the DMCA online safe harbor, the E-commerce Directive expects services to follow a notice-and-takedown scheme, i.e., services must remove or disable access to content in response to takedown notices.[54] However, while the DMCA online safe harbor only applied to alleged copyright infringement, the E-Commerce Directive required removals for all categories of illegal or tortious material.[55]

European countries have adapted the E-Commerce Directive's notice-and-takedown model for specific contexts. For example, in 2017, Germany

---

provider may prefer to disable access to material, rather than removing it, including so that a link may be restored in response to a counter notification or a court order in a lawsuit between the copyright owner and poster or to preserve evidence." IAN C. BALLON, 4 E-COMMERCE AND INTERNET LAW 4.12[6][C] (2020 update), Westlaw ECOMMINTLAW; *see* Rosen v. eBay, Inc., No. CV 13–6801 MWF (Ex), 2015 WL 1600081, at *11–12 (C.D. Cal. Jan. 16, 2015) (holding that eBay properly disabled access to files even if the URL still could be accessed by someone who knew the URL before it had been disabled).

47. 17 U.S.C. §§ 512(c)(1)(A)(iii) & 512(c)(1)(C).

48. *Id.* § 512(i)(1)(A).

49. *E.g.*, Ventura Content, Ltd. v. Motherless, Inc., 885 F.3d 597, 613–19 (9th Cir. 2018).

50. Shoshana Wodinsky, *YouTube's Copyright Strikes Have Become a Tool for Extortion*, VERGE (Feb. 11, 2019, 8:20 AM), https://www.theverge.com/2019/2/11/18220032/youtube-copystrike-blackmail-three-strikes-copyright-violation.

51. Many services have (or had) a three-strikes-and-you're-out policy, including YouTube and Tumblr. *See* Melanie Ehrenkranz, *YouTube Updates Its Three-Strikes Policy—But Not the One You're Mad About*, GIZMODO (Feb. 19, 2019, 1:51 PM), https://gizmodo.com/youtube-updates-its-three-strikes-policy-but-not-the-on-1832726224; Jonathan Bailey, *Don't Blame the DMCA for Tumblr's Policy*, PLAGIARISM TODAY (June 23, 2015), https://www.plagiarismtoday.com/2015/06/23/dont-blame-the-dmca-for-tumblrs-policy; *The Digital Millennium Copyright Act (DMCA)*, GIGANEWS, https://www.giganews.com/legal/dmca.html (last visited Oct. 22, 2021).

52. *See* BMG Rights Mgmt. (US) LLC v. Cox Commc'ns, Inc., 881 F.3d 293, 303 (4th Cir. 2018) (indicating that a 13-strike policy was too lax to retain the safe harbor).

53. Directive 2000/31, of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market, 2000 O.J. (L 178) 1.

54. *Id.* art. 14(1)(b).

55. Goldman, *Section 230 Overview*, *supra* note 5, at 167–68.

passed the Netzwerkdurchsetzungsgesetz, the Network Enforcement Act ("NetzDG"). NetzDG requires that services "remove or block access" to enumerated categories of illegal content within very short timeframes.[56] Similarly, the U.K. Defamation Act requires Internet services to remove allegedly defamatory user statements within 48 hours of a takedown notice unless the service provides the user's identifying information to the complainant.[57]

3. The Manila Principles

The prior two examples involved legal regulations encoding the binary approach to remedies. The next two examples come from statements issued by civil society organizations.

The Manila Principles on Intermediary Liability[58] are designed to guide "policymakers and intermediaries when developing, adopting, and reviewing legislation, policies and practices that govern the liability of intermediaries for third-party content."[59] In general, the Manila Principles promote free expression by discouraging governments from unreasonably suppressing user content.[60]

Given this objective, not surprisingly, the Manila Principles focus on content removals. One principle says: "Laws and content restriction orders and practices must comply with the tests of necessity and proportionality,"[61] including:

- "courts should only order the removal of the bare minimum of content that is necessary to remedy the harm identified;"[62]
- companies should adopt "the least restrictive technical means" of restricting content;[63]

56. *See generally* HEIDI TWOREK & PADDY LEERSSEN, TRANSATLANTIC WORKING GRP., AN ANALYSIS OF GERMANY'S NETZDG LAW (2019), https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf.

57. The Defamation (Operators of Websites) Regulations 2013, SI 2013/3028 (Eng. & Wales). The U.K. adopted the Defamation Act when it was part of the European Union and thus obligated to follow E.U. directives.

58. MANILA PRINCIPLES ON INTERMEDIARY LIABILITY, VERSION 1.0 (Mar. 24, 2015), https://www.eff.org/files/2015/10/31/manila_principles_1.0.pdf [hereinafter MANILA PRINCIPLES].

59. *Id.* at 1.

60. *See id.*; *cf.* GLOBAL NETWORK INITIATIVE, IMPLEMENTATION GUIDELINES FOR THE PRINCIPLES ON FREEDOM OF EXPRESSION AND PRIVACY para. 3.2 (2017), https://globalnetworkinitiative.org/wp-content/uploads/2018/08/Implementation-Guidelines-for-the-GNI-Principles.pdf (providing a similar approach for companies responding to government demands for content removal).

61. MANILA PRINCIPLES, *supra* note 58, at 4.

62. *Id.* at 35.

63. *Id.* at 36.

- companies should deploy geographically variegated content restrictions, so that restrictions are as geographically limited as possible;[64] and
- companies should deploy the most temporally limited content restrictions.[65]

The Manila Principles sometimes use the term "content restrictions" instead of "content removals," but the Manila Principles overwhelmingly focus on removals. For example, four of the five examples describing "content restrictions" explicitly relate to content removals or takedowns.[66]

4. Santa Clara Principles

In 2018, some civil society organizations and academics issued the Santa Clara Principles on Transparency and Accountability in Content Moderation.[67] The principles describe procedural due process approaches that Internet services should voluntarily adopt, including: what good transparency reports contain; how companies should provide detailed notices to users when taking actions; and the availability of user appeals for those actions. The principles explicitly discuss content removals and account suspensions.

5. The "Internet Balancing Formula"

The next example involves an academic proposal. In 2019, European law professor Mart Susi proposed an "Internet Balancing Formula" to balance the free expression value of content against reasons to suppress the content, such as privacy interests.[68] It assigns numerical values to various factors, some in favor of free expression and others in favor of content suppression, and computes a precise fraction of the factors.[69] For fractions less than one, the content should not be restricted because its free expression value predominates; if greater than one, the content "should not be published or should be blocked."[70]

---

64. *Id.* at 39.
65. *Id.* at 40.
66. *Id.* at 16–17. The fifth example is "notice and notice," where a service forwards a takedown notice to the targeted content uploader but otherwise takes no action. *Id.* at 17. *See* Copyright Act, R.S.C. 1985, c. C-42, § 41.26 (Can.).
67. *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, SANTA CLARA PRINCIPLES, https://santaclaraprinciples.org (last visited Oct. 22, 2021).
68. Mart Susi, *The Internet Balancing Formula*, 25 EUR. L.J. 198 (2019) [hereinafter Susi*, Balancing*]; *see also* Robert Alexy, *Mart Susi's Internet Balancing Formula*, 25 EUR. L.J. 213 (2019); Mart Susi, *Reply to Robert Alexy's Critique of the Internet Balancing Formula*, 25 EUR. L.J. 221 (2019).
69. Susi, *Balancing*, *supra* note 68, at 204–07.
70. *Id.* at 207.

This formula operationalizes the E.U. E-Commerce Directive, which necessitates that the formula treats removal as the only applicable remedy.[71] Yet, the formula seems tailor-made for implementing alternative remedies in close cases. For example, if the formula yielded a result between 0.5 and 1.5, the closeness of the question might warrant some intervention other than removal. Part IV(A)(2) will address the relevance of close decisions when deciding the appropriate remedies.

6. Principles for User Generated Content Services

As the prior five examples indicate, content removal and account termination are widely incorporated into the content moderation discourse. The next two examples differ from the prior five because they expressly incorporated alternative remedy schemes.

In 2007, some copyright owners announced their "Principles for User Generated Content Services."[72] These principles sought to induce "services providing user-uploaded and user-generated audio and video content" to work harder to prevent user-caused copyright infringement.[73] Copyright owner signatories agreed not to sue Internet service signatories for copyright infringement if the services satisfied the principles' very exacting requirements.[74] Those requirements included blocking users' uploads that matched a database of precedent works, unless the copyright owner "wishes to exercise an alternative to blocking (such as allowing the content to be uploaded, licensing use of the content or other options)."[75]

Unfortunately, the principles did not elaborate on these blocking alternatives. The principles appear to contemplate YouTube's Content ID program, which allows copyright owners to acquiesce to user-uploaded works that copy their material and claim any generated revenues.[76] The principles ultimately fizzled out due to Internet services' lack of enthusiasm for the weak benefits.[77]

---

71. Directive 2000/31, of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market, art. 14, § 1(b), 2000 O.J. (L 178) 1, 13.

72. *Principles for User Generated Content Services*, UGC PRINCIPLES, http://ugcprinciples.com (last visited Oct. 22, 2021); *see* Note, *The Principles for User Generated Content Services: A Middle-Ground Approach to Cyber-Governance*, 121 HARV. L. REV. 1387 (2008).

73. *Principles for User Generated Content Services*, *supra* note 72.

74. *Id.* para. 14.

75. *Id.* para. 3(c).

76. *How Content ID Works*, YOUTUBE HELP, https://support.google.com/youtube/answer/2797370?hl=en (last visited Sept. 16, 2021).

77. In particular, one signatory, Veoh, qualified for the DMCA online safe harbor. UMG Recordings, Inc. v. Shelter Cap. Partners LLC, 667 F.3d 1022, 1030–45 (9th Cir. 2011). Nevertheless, copyright owners sued it into bankruptcy. Eric Goldman, UMG v. Shelter Capital*: A Cautionary Tale of Rightsowner Overzealousness*, TECH. & MKTG. L. BLOG (Dec. 20, 2011), https://blog.ericgoldman.org/archives/2011/12/umg_v_shelter_c.htm.

7. Graduated Response/Copyright Alert System

In the late 2000s, copyright owners wanted Internet access providers (IAPs) to discourage copyright infringement by their subscribers. This led to a solution called "Graduated Response," which imposed escalating consequences on IAP subscribers who repeatedly used file-sharing software to infringe.[78]

IAPs differ from other Internet services, such as web hosts or social media services, in important ways. First, IAPs cannot control individual content items disseminated by subscribers (except by using disfavored techniques like deep-packet inspection[79]), so IAPs have fewer remedy options. Second, restrictions on Internet access may interfere with the subscriber's ability to use the Internet at all—a potentially life-altering and disproportionate penalty.[80] Still, the graduated response initiatives have prompted some interesting remedies experiments at IAPs.

Graduated Response (Riposte Graduée) in France

France adopted a graduated response program called "HADOPI," named for the government agency charged with its enforcement.[81] It is commonly called the "Three Strikes" law due to the number of infringement claims before the IAP subscriber experiences serious consequences.[82] The remedies first attempt to educate users and then impose harsher remedies on recidivists, as follows:

- Strike 1: email warning.[83]
- Strike 2: warning sent in the postal mail.
- Strike 3: the subscriber is referred to court, which can impose a fine of up to 1,500. Prior to 2013, the court also could

78. *E.g.*, Peter K. Yu, *The Graduated Response*, 62 FLA. L. REV. 1373, 1374 (2010).

79. *See* Catherine J.K. Sandoval, *Disclosure, Deception, and Deep-Packet Inspection: The Role of the Federal Trade Commission Act's Deceptive Conduct Prohibitions in the Net Neutrality Debate*, 78 FORDHAM L. REV. 641, 646 (2009); Annemarie Bridy, *Graduated Response American Style: 'Six Strikes' Measured Against Five Norms*, 23 FORDHAM I.P., MEDIA & ENTER. L.J. 1, 44–46 (2012) [hereinafter Bridy*, American Style*].

80. *E.g.*, Frank La Rue (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 85, U.N. Doc. A/HRC/17/27 (May 16, 2011) ("[T]he Internet has become an indispensable tool for realizing a range of human rights, combating inequality, and accelerating development and human progress . . . .").

81. The agency is "Haute Autorité pour la Diffusion des Oeuvres et la Protection des droits d'auteur sur Internet."

82. *See* Sandrine Rambaud, *Illegal Internet File Downloads Under HADOPI 1 and 2*, 15 CYBERSPACE LAW. 10 (2010).

83. Starting in 2015, the protocol added a "reminder" letter sent by mail to supplement the email warning. HADOPI, 2016/17 ACTIVITY REPORT 24, https://www.hadopi.fr/sites/default/files/sites/default/files/ckeditor_files/Activity-report-2016-17-HADOPI.pdf.

> temporarily suspend Internet access, and the subscriber would be blocklisted from obtaining services from other IAPs.[84]

There is widespread skepticism about HADOPI's cost-benefit.[85]

The Copyright Alert System

The U.S. Congress has not adopted a graduated response statutory requirement,[86] but in 2011, copyright owners promulgated a "voluntary"[87] program called the "Copyright Alert System."[88]

Like HADOPI, the Copyright Alert System imposed escalating remedies for users' alleged infringement by file-sharing.[89] The first few strikes triggered educational warnings to allegedly infringing subscribers. After further recidivism, the IAP then implemented "mitigation measures" such as:

---

84. Andy Maxwell, *Three Strikes and You're Still In – France Kills Piracy Disconnections*, TORRENTFREAK (July 9, 2013), https://torrentfreak.com/three-strikes-and-youre-still-in-france-kills-piracy-disconnections-130709.

85. Glyn Moody, *In 10 Years Of Existence, The Long-Running French Farce Known As Hadopi Has Imposed Just €87,000 In Fines, But Cost Taxpayers €82 Million*, TECHDIRT (Aug. 6, 2020, 3:24 AM), https://www.techdirt.com/articles/20200805/07053345044/10-years-existence-long-running-french-farce-known-as-hadopi-has-imposed-just-87000-fines-cost-taxpayers-82-million.shtml.

86. *See* Annemarie Bridy, *Graduated Response and the Turn to Private Ordering in Online Copyright Enforcement*, 89 OR. L. REV. 81, 81–82 (2010).

87. Similar to the Principles for User Generated Content Services, copyright owners encouraged IAPs to voluntarily participate in the Copyright Alert System to avoid ruinous copyright infringement litigation. That was not an empty threat. Since the Copyright Alert System's demise, copyright owners have sued IAPs for subscribers' purported infringements. *E.g.*, BMG Rights Mgmt. (US) LLC v. Cox Commc'ns, Inc., 881 F.3d 293 (4th Cir. 2018); UMG Recordings, Inc. v. Grande Commc'ns Networks, LLC, 384 F. Supp. 3d 743 (W.D. Tex. 2019); Warner Records Inc. v. Charter Commc'ns, Inc., No. 19-cv-00874-RBJ-MEH, 2020 WL 6511988 (D. Colo. Nov. 5, 2020); UMG Recordings, Inc. v. Bright House Networks, LLC, No. 8:19-cv-710-MSS-TGW, 2020 WL 3957675 (M.D. Fla. July 8, 2020); ALS Scan, Inc. v. Steadfast Networks LLC, 819 F. App'x. 522 (9th Cir. 2020); Sony Music Ent. v. Cox Commc'ns, Inc., 426 F. Supp. 3d 217 (E.D. Va. 2019); *see* Eric Goldman, *Internet Access Provider May Be Vicariously Liable for Subscribers' BitTorrent Downloads*–Warner Bros. v. Charter, TECH. & MKTG. L. BLOG (Oct. 28, 2019), https://blog.ericgoldman.org/archives/2019/10/internet-access-provider-may-be-vicariously-liable-for-subscribers-bittorrent-downloads-warner-bros-v-charter.htm.

88. PUBLIC INTELLIGENCE, MEMORANDUM OF UNDERSTANDING CREATING THE CENTER FOR COPYRIGHT INFORMATION (2011), https://info.publicintelligence.net/CCI-MOU.pdf [hereinafter CAS MOU]. *See generally* Corynne McSherry & Eric Goldman, *The "Graduated Response" Deal: What if Users Had Been at the Table?*, ELEC. FRONTIER FOUND. (July 18, 2011), https://www.eff.org/deeplinks/2011/07/graduated-response-deal-what-if-users-had-been.

Because the initiative specified consequences for six incidents of claimed infringement by a subscriber, it was sometimes called the "six strikes" program. *E.g.*, Karl Bode, *'Six Strikes' May Be Dead, But ISPs Keep Threatening to Disconnect Accused Pirates Anyway*, TECHDIRT (Oct. 4, 2017, 6:20 AM), https://www.techdirt.com/articles/20171003/09553238335/six-strikes-may-be-dead-isps-keep-threatening-to-disconnect-accused-pirates-anyway.shtml.

89. CAS MOU, *supra* note 88, para. 4(G). *See generally* Bridy, *American Style*, *supra* note 79, at 30–37.

- temporarily reduce upload/download speeds;
- reduce the subscriber's service tier to "(1) the lowest tier of Internet access service above dial-up service that the Participating ISP makes widely available to residential customers in the Subscriber's community, or (2) an alternative bandwidth throughput rate low enough to significantly impact a Subscriber's broadband Internet access service (e.g., 256 - 640 kbps);"
- "temporary redirection to a Landing Page until the Subscriber contacts the Participating ISP to discuss with it the Copyright Alerts;"
- "temporary restriction of the Subscriber's Internet access for some reasonable period of time as determined in the Participating ISP's discretion;"
- "temporary redirection to a Landing Page for completion of a meaningful educational instruction on copyright."[90]

The Copyright Alert System gave IAPs some discretion about which mitigation measures to implement. Shortly following the launch, IAPs chose different options as their most severe remedy[91]:

Company	"Harshest" Remedy
Comcast	View mandatory video and in-browser alert
Verizon	Reduce transmission speed
Time Warner Cable	Account suspended until user calls in and apologizes
AT&T	Account suspended until user completes an IP course
Cablevision	Up to 48 hours of account suspension

The Copyright Alert System shut down after four years of operation,[92] though IAPs still may voluntarily deploy some or all of its contemplated remedies.[93]

---

90. CAS MOU, *supra* note 88, at para. 4(G)(iii). However, the IAP was not required to implement a measure that "knowingly disables or is reasonably likely to disable a Subscriber's access to any IP voice service (including over-the-top IP voice service), e-mail account, or any security service, multichannel video programming distribution service or guide, or health service (such as home security or medical monitoring) while a Mitigation Measure is in effect." *Id*. *See generally* Bridy, *American Style*, *supra* note 79, at 31–33.

91. *See* Rebecca Greenfield, *You Will Be Warned: ISPs Roll Out Their Anti-Piracy Alert Systems*, ATLANTIC (Feb. 26, 2013), https://www.theatlantic.com/technology/archive/2013/02/you-will-be-warned-isps-roll-out-their-anti-piracy-alert-systems/317977; Bryan Bishop, *Comcast and Cablevision Detail Their 'Six Strike' Copyright Alert Strategies*, VERGE (Feb. 27, 2013, 10:22 PM), https://www.theverge.com/2013/2/27/4038184/comcast-and-cablevision-detail-their-six-strike-copyright-alert-strategies.

92. *E.g.*, David Kravets, *RIP, "Six Strikes" Copyright Alert System*, ARS TECHNICA (Jan. 30, 2017, 2:50 PM), https://arstechnica.com/tech-policy/2017/01/rip-six-strikes-copyright-alert-system.

Why Didn't Alternative Remedies Work?

The copyright-related experiments with alternative remedies, including the Principles for User Generated Content and graduated response initiatives, have not achieved the copyright owners' objectives. This is not surprising, nor does it predict the potential success of alternative remedial schemes more generally. The copyright owner constituency has sought to interject its desired remedies into the Internet service/user relationships. Indeed, the Copyright Alert System put the IAPs into positions adverse to their paying subscriber-customers. The remedial schemes were not designed to advance the interests of the service or its users, and that undermined their likely efficacy. This should caution regulators about the risks of mandating specific remedies—especially if the remedies are intended to benefit a self-interested lobby.

B. *Moving Beyond Removals*

As the prior subpart demonstrated, regulators and commentators have historically treated the removal remedy as the paramount solution for violative content or actions. It is easy to imagine how removals emerged as the "default" remedy for redressing legal violations. Regulators can easily describe the remedy; Internet services universally can comply with it (more complex remedies may require custom programming or may not be functionally possible for certain services); removals prevent ongoing legal violations; and removals are easily measured and verified. The late 1990s' adoption of the DMCA and the E.U. E-Commerce Directive did much to shape global regulatory norms, and at that time, the risks and consequences of over-removals were less obvious to regulators.[94] By the time those consequences became more widely recognized, the global regulatory norms in support of the removal remedy were ingrained.[95]

---

93. Andy Maxwell, *Six Strikes Piracy Scheme May Be Dead But Those Warnings Keep on Coming*, TORRENTFREAK (Oct. 1, 2017), https://torrentfreak.com/six-strikes-piracy-scheme-may-be-dead-but-those-warnings-keep-on-coming-171001.

94. *See generally* Klonick, *supra* note 19 (providing historical background on Internet services' chaotic and unsystematic development of their approaches to content moderation).

Interestingly, Congress' original attempt at Internet regulation, the Communications Decency Act (the CDA), essentially sought to push commercial pornography behind an age-authenticated registration wall rather than remove it outright. Communications Decency Act of 1996, Pub. L. No. 104-104, 110 Stat. 133 (1996). However, due to terrible statutory design, the law would have functionally required removal of the targeted content; and the technological infrastructure to implement the metaphorical age-gate did not exist at the time and would have been cost-prohibitive for many services. Reno v. ACLU, 521 U.S. 844, 847 (1997).

95. As an indicator that "removals" are deeply entrenched into corporate architecture, Google's content moderation function includes a "Legal Removals" team. *See* Gareth Corfield, *Here is How Google Handles Right to Be Forgotten Requests*, REGISTER (Mar. 19, 2018, 9:43 AM), https://www.theregister.com/2018/03/19/google_right_to_be_forgotten_request_process.

Unfortunately, this regulatory "obsession with removal"[96] has hindered the consideration of other remedial options. Expanding the remedies beyond removals carries several benefits.

First, removals can cause collateral damage.[97] Tarleton Gillespie explained that the removals remedy "is the harshest approach, in terms of its consequences . . . . Removal is a blunt instrument, an all-or-nothing determination."[98] Some problems that removals may cause:

- Removals wipe away evidence of the violation, leaving a hole in the community's historical record. For example, when Twitter suspended President Trump's account, it depublished all of Trump's tweets despite their critical importance to the historical record.[99] Evidence also suggests that some victims of online harassment are harmed when the harassing content is removed because it hides the evidence of the anti-social behavior they suffered.[100]
- When a service deletes a content item, it must either delete any comments that are part of the same thread, or leave those comments orphaned and decontextualized (like what happened to all of the tweeted responses to President Trump's depublished tweets).
- Similarly, removals break inbound links, which degrades the user experience for anyone following the links.
- In the case of account removals, the collateral damage includes: (1) the removal of any non-violative content associated with that account, (2) restricted usage of other services offered by the same company (which can be a problem for diversified enterprises like Google and Facebook),[101] and (3) difficulty logging into third-party services that have linked their account authorizations.[102]

---

96. Common, *supra* note 16, at 135; *see also* Daphne Keller, *Empirical Evidence of "Over-Removal" by Internet Companies Under Intermediary Liability Laws*, STAN. L. SCH. CTR. FOR INTERNET & SOC'Y BLOG (Oct. 12, 2015, 8:23 AM), http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws.

97. GILLESPIE, *supra* note 19, at 176.

98. *Id*.

99. *See* David Gewirtz, *Why All of Trump's Tweets and Other Social Media Posts Must Be Archived for Future Historians*, ZDNET (Jan. 12, 2021), https://www.zdnet.com/article/why-all-of-trumps-tweets-and-other-social-media-posts-must-be-archived-for-future-historians.

100. Schoenebeck et al., *supra* note 31, 1292–96.

101. For example, terminating a Facebook account excludes the accountholder from Facebook communities that exist nowhere else. Or termination of a Zoom account might effectively expel a student from an online-only school.

102. See OAUTH, https://oauth.net (last visited Oct. 17, 2021). *See generally* Kashmir Hill, *I Tried to Live Without the Tech Giants. It Was Impossible.*, N.Y. TIMES (July 31, 2020), https://www.nytimes.com/2020/07/31/technology/blocking-the-tech-giants.html (discussing

While removals will always be an integral part of Internet services' remedial toolkit,[103] they should not be the only tool[104]—and perhaps not even the most important tool.[105] A velvet glove works better than a sledgehammer in some circumstances.[106] An expanded remedy tool kit allows for more tailored and nuanced outcomes that can balance the benefits and harms from continued publication.[107] This advances free expression while still redressing content that violates service providers' content rules.[108]

Second, expanded non-removal remedies may alleviate the widely held perception that Internet services "censor" its users.[109] Despite the fact that the Constitution only restricts government "censorship,"[110] some users nevertheless feel "censored" by Internet services when their accounts or content are removed.[111] Those feelings of censorship have contributed to animus towards "Big Tech," which has fueled demands for legal reforms

---

what happened when the reporter tried to avoid using services like Google and Amazon, only to realize how many other Internet services are linked to them).

103. YouTube Report, *supra* note 22, at 3 ("[R]emoval of content is an important lever we use to address information quality.").

104. Douek, *supra* note 33, at 787–88; YouTube Report, *supra* note 22, at 3 (stating removal "is not the only lever at our disposal, and we use it with caution").

105. "Deleting content is not a solution; it is simply a 'Band-Aid' for an already existing problem." BEN WAGNER ET AL., REIMAGINING CONTENT MODERATION AND SAFEGUARDING FUNDAMENTAL RIGHTS: A STUDY ON COMMUNITY-LED PLATFORMS 29 (2021). According to content moderation expert Alex Feerst, "[R]emoval happens because subtler and more constructive solutions have failed or don't exist." Email from Alex Feerst to Eric Goldman (Jan. 21, 2021) (on file with author). *See also* WAGNER ET AL., *supra*, at 16, 18 (describing how deletion is considered a remedy of last resort on services like diaspora* and Mastadon).

106. As Gillespie described it, "removing content or users is akin to the most profound kind of censorship." GILLESPIE*, supra* note 19, at 177.

107. *Cf.* LEE ANNE FENNELL, SLICES AND LUMPS: DIVISION AND AGGREGATION IN LAW AND LIFE (2019) (discussing how divisible remedies can help blunt the effects of indivisible laws); Adam J. Kolber, *Smooth and Bumpy Laws*, 102 CALIF. L. REV. 655 (2014) (discussing the disadvantages of disproportionate consequences from legal violations); Douek, *supra* note 33 (discussing the problems with categorical rule-based content moderation).

108. *See* Molly K. Land & Rebecca J. Hamilton, *Beyond Takedown*: *Expanding the Toolkit for Responding to Online Hate*, *in* PROPAGANDA, WAR CRIMES TRIALS AND INTERNATIONAL LAW: FROM COGNITION TO CRIMINALITY 143 (Predrag Dojcinovic ed., 2020).

109. *E.g.*, Emily A. Vogels, Andrew Perrin & Monica Anderson, *Most Americans Think Social Media Sites Censor Political Viewpoints*, PEW RSCH. CTR. (Aug. 19, 2020), https://www.pewresearch.org/internet/2020/08/19/most-americans-think-social-media-sites-censor-political-viewpoints.

110. *E.g.*, Manhattan Cmty. Access Corp. v. Halleck, 587 139 S. Ct 1921, 1926–33 (2019).

111. Numerous users have sued Internet services (often pro se) for "censoring" them. *E.g.*, Divino Group LLC v. Google LLC, No. 19-cv-04749-VKD, 2021 WL 51715 (N.D. Cal. Jan. 6, 2021); Elansari v. Jagex Inc., 790 F. App'x. 488 (3d Cir. 2020); Belknap v. Alphabet, Inc., 504 F. Supp. 3d 1156 (D. Ore. 2020); Lewis v. Google LLC, 461 F. Supp. 3d 938 (N.D. Cal. 2020); Shulman v. Facebook.com, No. 17–764 (JMV) (LDW), 2018 WL 3344236 (D. N.J. July 9, 2018); *see* Goldman & Miers, *supra* note 21, at 196–204, 217–20.

such as the amendment or repeal of Section 230.[112] By creating better balances between free expression and content policing, wider deployment of non-removal remedies may reduce public ire and the temperature of the policy debates.

Third, online communities have diverse audiences with idiosyncratic needs. If removals are the exclusive or primary remedy across all services, then the services will lose some of their distinctive natures. In contrast, an expanded remedy toolkit will let Internet services refine and optimize their content moderation approaches to best cater to their specific community's needs.[113] Indeed, a service's remedy "strategy" can become a key point of competitive differentiation.[114] Services competing for the same audiences can adopt differing strategies and let audiences decide which approach creates the kind of community or resources they want.[115] Thus, an expanded remedy toolkit beyond removals can enhance marketplace competition and help services do a better job catering to their audiences.

III. A TAXONOMY OF REMEDY OPTIONS

This Part enumerates about three dozen remedy options for violations of online rules. None of these options are hypothetical or conjectural; all have been deployed by at least one service. From a technology standpoint, the range of potential remedies is essentially infinite—and with Section 230's immunity,[116] that may also be true from a legal standpoint (when regulators do not otherwise mandate particular remedies).[117]

The remedy taxonomy has five categories:

(1) actions against individual content items;

---

112. *See* Eric Goldman, *While Our Country Is Engulfed By Urgent Must-Solve Problems, Congress Is Working Hard to Burn Down Section 230*, TECH. & MKTG. L. BLOG (Aug. 4, 2020), https://blog.ericgoldman.org/archives/2020/08/while-our-country-is-engulfed-by-urgent-must-solve-problems-congress-is-working-hard-to-burn-down-section-230.htm.

113. *See* Land & Hamilton, *supra* note 108.

114. *See* Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INST. (Feb. 11, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels (raising concerns about cross-industry "cartels" that establish uniform content policies across the industry); *cf.* YouTube Report, *supra* note 22, at 6 (stating that house rules "represent a crucial part of what makes that product unique").

115. With a minor caveat that diverse remedial schemes might inhibit users' willingness to migrate to new services because they will have to learn new remedial schemes.

116. In general, Section 230 sought to minimize regulatory impact on Internet services' editorial decisions, including the decisions about which remedies to deploy. Section 230 contained a finding that the "Internet and other interactive computer services have flourished, to the benefit of all Americans, with a minimum of government regulation." 47 U.S.C. § 230(a)(4). Section 230 also stated a policy objective "to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation." *Id.* § 230(b)(2).

117. Eric Goldman, *Internet Immunity and the Freedom to Code*, 62 COMM. ACM 22 (2019).

(2) actions against an online account;

(3) actions to reduce the visibility of violations, which can be implemented against individual content items or an entire account;

(4) actions to impose financial consequences for violations, which also can be implemented against individual content items or an entire account; and

(5) a miscellaneous category for actions that do not fit into the other categories.[118]

This chart summarizes the taxonomy and remedy options[119]:

Content Regulation	Account Regulation	Visibility Reductions (by acct or item)	Monetary (by acct or item)	Other
• Remove content • Suspend content • Relocate content • Edit/redact content • Interstitial warning • Add warning legend • Add counterspeech • Disable comments	• Terminate account • Suspend account • Suspend posting rights • Remove credibility badges • Reduced service levels (data, speed, etc.) • Shaming	• Shadowban • Remove from external search index • Nofollow authors' links • Remove from internal search index • Downgrade internal search visibility • No auto-suggest • No/reduced internal promotion • No/reduced navigation links • Reduced virality • Age-gate • Display only to logged-in readers	• Forfeit accrued earnings • Terminate future earning (by item or account) • Suspend future earning (by item or account) • Fine author/impose liquidated damages	• Educate users • Assign strikes/warnings • Outing/unmasking • Report to law enforcement • Put user/content on blocklist • Community service • "Restorative justice"/apology

More detailed descriptions of each remedy:

---

118. Kraut & Resnick and Grimmelmann have previously offered related taxonomies. Kraut & Resnick's taxonomy included: (1) selection, sorting, highlighting, (2) community structure, (3) feedback and rewards, (4) access controls, (5) roles, rules, policies, and procedures, and (6) presentation & framing. ROBERT E. KRAUT & PAUL RESNICK, BUILDING SUCCESSFUL ONLINE COMMUNITIES: EVIDENCE-BASED SOCIAL DESIGN 168–69 tbl.4.1 (Douglas Sery & Mel Goldsipe eds., 2012). Grimmelmann summarized his four-node taxonomy: "Exclusion keeps unwanted members out of the community entirely; pricing uses market forces to allocate participation. . . . In organization, moderators reshape the flow of content from authors to readers; in norm-setting, they inculcate community-serving values in other members." Grimmelmann, *supra* note 12, at 55. The Kraut & Resnick and Grimmelmann taxonomies both combine pre-violation content moderation efforts with post-violation remedies. This Article only taxonomizes post-violation remedies.

Google/YouTube adopted a "4 R" taxonomy: remove, raise, reduce, and reward. YouTube Report, *supra* note 22, at 4.

119. For a similar chart that includes the pros/cons of options, see WAGNER ET AL., *supra* note 105, at 22–23 tbl.1.

A. *Content Regulation*[120]

This subpart describes eight remedies against individual content items.

1. *Remove Content*: Permanently remove content. This can be done once ("takedown") or as an ongoing ban of the content (a "staydown" remedy).[121] Removals can be made on a network-wide global basis or only in specific geographies or parts of the network. Per Part II, this is the "standard" remedy.

2. *Suspend Content*: Remove content temporarily—from anywhere between minutes and forever. Indefinite content suspensions are functionally equivalent to content removal. Internet services routinely suspend content.

Examples:

- Medium suspends controversial, suspect, and extreme content.[122]
- WordPress suspends content that violates its policies.[123]

3. *Relocate Content*: A content item gets deleted at its current URL and uploaded to a new URL. This change in URLs resets the number of user views, removes user comments, and breaks inbound links. These consequences may frustrate the uploader's promotional efforts.

Example: YouTube relocates videos it believes are promoted by spam.[124]

4. *Edit/Redact Content*: Instead of removing an item entirely, a service can edit out or redact only the violative portion. This approach may undermine the service's eligibility for Section 230 immunity when the edits create the tortious or illegal aspects.[125] However, Section 230 may protect editing illegal or tortious content to make it legal.[126]

---

120. *Cf.* Grimmelmann, *supra* note 12, at 58–61 (discussing "organizing" content).

121. Martin Husovec, *The Promises of Algorithmic Copyright Enforcement: Takedown or Staydown? Which Is Superior? And Why?*, 42 COLUM. J.L. & ARTS 53, 57 (2018) ("Depending on the scope of preventive 'staydown' obligation, it might require an intermediary to protect from re-infringing only (1) in the same form (e.g. re-uploading of an identical file with a full copyrighted work), or (2) in any other form (e.g. re-uploading a part of the work)."). Sometimes, staydown is called "notice-and-staydown" as an allusion to the "notice-and-takedown" phrase.

122. *Controversial, Suspect, and Extreme Content,* MEDIUM, https://help.medium.com/hc/en-us/articles/360018182453 (last visited Oct. 23, 2021).

123. *Suspended Content and Sites,* WORDPRESS, https://en.support.wordpress.com/suspended-blogs (last visited Oct. 23, 2021).

124. *See*, *e.g.*, Kinney v. YouTube, LLC, No. G054863, 2018 WL 5961898 (Cal. Ct. App. Nov. 14, 2018); Song Fi v. Google, Inc., No. 14-cv-05080-CW, 2018 WL 2215836 (N.D. Cal. May 15, 2018); Bartholomew v. YouTube, LLC, 225 Cal. Rptr. 3d 917 (Ct. App. 2017).

125. Fair Hous. Council of San Fernando Valley v. Roommates.com, LLC, 521 F.3d 1157, 1162–63 (9th Cir. 2008).

126. *See* Court's Final Ruling on Demurrer at 9, People v. Ferrer, No. 16FE019224 (Cal. Super. Ct. Dec. 9, 2016), http://digitalcommons.law.scu.edu/cgi

Examples:

- Ripoff Report removes statements in user-provided reviews that its private arbitration service determines are defamatory.[127]
- Backpage allegedly edited user-submitted prostitution ads to remove indicia of illegal behavior.[128]
- Discourse.net may "disemvowel" "comments that are duplicative, commercial, needlessly foul or mean or otherwise inappropriately offensive."[129] "Disemvoweling" means to remove all of the vowels from violative content.[130]

5. *Interstitial Warning*: Interpose a warning before readers access the content.

Examples:

- Facebook imposes an interstitial warning on "graphic" photos and videos.[131]
- Twitter places a notice of violation on tweets that it leaves up as being in the public interest.[132]

/viewcontent.cgi?article=2358&context=historical ("[T]he People are essentially complaining that Backpage staff scrubbed the original ad, removing any hint of illegality. If this was the alleged content 'manipulation,' the content was modified from being illegal to legal.") (citation omitted). Other courts have reached contrary conclusions. *See* United States v. Lacey, 423 F. Supp. 3d 748, 759–60 (D. Ariz. 2019); United States v. $1,546,076.35 In Bank Funds Seized From Republic Bank of Arizona Account 1889, No. 2:18-cv-08420-RGK-PJW, 2020 WL 8172984, at *8–10 (C.D. Cal. Dec. 1, 2020).

127. *VIP Arbitration Program*, *supra* note 24 ("Even if the Arbitrator determines that a Report contains one or more false statements of fact, this does not mean the whole Report will be removed. Instead, any Report and associated Comment thereto found to be 'substantially false' by the Arbitrator will be redacted and replaced . . . .").

128. Court's Final Ruling on Demurrer at 5, People v. Ferrer (2016) (No. 16FE019224).

129. *Comment Policy*, DISCOURSE.NET (Apr. 29, 2009), https://www.discourse.net /comment-policy.

130. Cory Doctorow, *How to Keep Hostile Jerks from Taking over Your Online Community*, INFORMATIONWEEK (May 15, 2007), http://www.informationweek.com/how-to-keep-hostile-jerks-from-taking-over-your-onlinecommunity/d/d-id/1055100; *Best Inventions of 2008*, *Disemvoweling*, TIME, http://content.time.com/time/specials/packages/article/0,28804,1852747_1854195_1854185,00.html (last visited Oct. 23, 2012). *See generally* Copia Inst., *BoingBoing Begins Disemvoweling the Trolls (2007)*, TR. & SAFETY FOUND. (Aug. 10, 2021), https://www.tsf.foundation/blog/boingboing-begins-disemvoweling-the-trolls-2007 (providing a case study of this remedial technique).

131. *Violence and Graphic Content,* FACEBOOK TRANSPARENCY CTR., https://www.facebook.com/communitystandards/graphic_violence (last visited Oct. 23, 2021); *Why Am I Seeing a Warning Before I Can View a Photo or Video?,* FACEBOOK HELP CTR., https://www.facebook.com/help/814083248683500 (last visited Oct. 23, 2021). Instagram has a similar approach for "sensitive content." *Why Am I Seeing a Warning Before I Can View a Photo or Video on Instagram?,* INSTAGRAM HELP CTR., https://www.facebook.com/help /instagram/188848648282410 (last visited Oct. 23, 2021).

132. *About Public-Interest Exceptions on Twitter*, TWITTER HELP CTR., https://help.twitter.com/en/rules-and-policies/public-interest (last visited Oct. 23, 2021) ("[W]e may choose to leave up a Tweet from an elected or government official that would

- Reddit requires users to affirmatively opt into "quarantined" subreddits.[133]
- YouTube imposes interstitial warnings on inflammatory religious or supremacist content.[134]

6. *Add Warning Legend*: Display a warning on the same screen as violative content.

Examples:

- Google adds warnings to unsafe search results.[135]
- TikTok adds a warning, "[t]he action in this video could result in serious injury," to videos depicting potentially dangerous stunts.[136]

7. *Add Counterspeech*: Place diverse or alternative perspectives next to content.[137]

Examples:

- Facebook adds links and snippets to "fact check" false stories.[138]
- Twitter displays a "Know the Facts" information bar above problematic or controversial topics like anti-vaccine content.[139]
- Tumblr displays public service announcements alongside search results for keywords related to eating disorders and self-harm.[140]

---

otherwise be taken down. Instead we will place it behind a notice providing context about the rule violation that allows people to click through to see the Tweet.").

133. *Quarantined Subreddits,* REDDIT HELP, https://www.reddithelp.com/en/categories/rules-reporting/account-and-community-restrictions/quarantined-subreddits (last visited Oct. 23, 2021).

134. Kent Walker, *Four Ways Google Will Help to Tackle Extremism*, FIN. TIMES (June 18, 2017), https://www.ft.com/content/ac7ef18c-52bb-11e7-a1f2-db19572361bb.

135. *Manage Warnings About Unsafe Sites,* GOOGLE CHROME HELP, https://support.google.com/chrome/answer/99020 (last visited Oct. 23, 2021).

136. Audra Schroder, *People Risking Concussions for New TikTok Challenge*, DAILY DOT (May 19, 2021, 10:32 PM), https://www.dailydot.com/unclick/tiktok-basketball-throwing-meme.

137. In a non-remedial context, Google News once allowed people quoted in news stories to provide additional context for their quotes. *Perspectives About the News from People in the News*, GOOGLE NEWS BLOG (Aug. 7, 2007), https://news.googleblog.com/2007/08/perspectives-about-news-from-people-in.html.

138. *Fact-Checking on Facebook: What Publishers Should Know*, FACEBOOK FOR BUS., https://www.facebook.com/help/publisher/182222309230722 (last visited Oct. 23, 2021); *see also Combatting Misinformation on Instagram*, FACEBOOK NEWSROOM (Dec. 16, 2019), https://about.fb.com/news/2019/12/combatting-misinformation-on-instagram.

139. Del Harvey, *Helping You Find Reliable Public Health Information on Twitter,* TWITTER BLOG (May 10, 2019), https://blog.twitter.com/en_us/topics/company/2019/helping-you-find-reliable-public-health-information-on-twitter.html.

8. *Disable Comments*: Disable additional user comments to user postings.

Examples:

- YouTube disables comments on videos that contain inflammatory religious or supremacist content.[141]
- Wikipedia editors can restrict users' ability to edit pages that are under attack (called page "protection").[142]

B. *Account Regulation*

This subpart describes six actions against a user's account.

1. *Terminate Account*: Permanently remove accounts. This is sometimes called "deplatforming."[143] As discussed in Part II, this is a standard remedy.

2. *Suspend Account*: Prevent users from accessing their accounts temporarily, ranging between minutes and forever. Permanent and indefinite suspensions are functionally indistinguishable from account termination.[144] Account suspensions are widely used.

Examples:

- Twitter suspends accounts that are "spammy" or "just plain fake."[145]
- Snapchat temporarily locks accounts for users engaged in prohibited activity.[146]

---

140. *A New Policy Against Self-Harm Blogs*, TUMBLR STAFF (Feb. 23, 2012), https://staff.tumblr.com/post/18132624829/self-harm-blogs.

141. Walker, *supra* note 134.

As a non-remedial example, YouTube also disables comments on videos featuring minors to thwart other users from sexualizing the depicted minor. *More Updates on Our Actions Related to the Safety of Minors on YouTube*, YOUTUBE OFF. BLOG, (Feb. 28, 2019), https://youtube-creators.googleblog.com/2019/02/more-updates-on-our-actions-related-to.html.

142. Wikipedia:Protection Policy, WIKIPEDIA, https://en.wikipedia.org/wiki/Wikipedia:Protection_policy (Oct. 22, 2021, 8:38 PM).

143. *E.g.*, Richard Rogers, *Deplatforming: Following Extreme Internet Celebrities to Telegram and Alternative Social Media*, 35 EUR. J. COMMC'N. 213, 214 (2020); Glenn Harlan Reynolds, *When Digital Platforms Become Censors*, WALL ST. J. (Aug. 18, 2018, 10:35 AM), https://www.wsj.com/articles/when-digital-platforms-become-censors-1534514122; *Is Deplatforming Enough To Fight Disinformation And Extremism?*, NPR (Jan. 25, 2021, 4:16 PM), https://www.npr.org/2021/01/25/960466075/is-deplatforming-enough-to-fight-disinformation-and-extremism.

144. *Cf. Case Decision 2021-001-FB-FBR*, OVERSIGHT BD. (May 5, 2021), https://www.oversightboard.com/sr/decision/2021/001/pdf-english [hereinafter *Oversight Bd. Decision*] (decision over Pres. Trump's Facebook account) (raising concerns about whether indefinite account suspensions are different from account terminations).

145. *About Suspended Accounts*, TWITTER HELP CTR., https://help.twitter.com/en/managing-your-account/suspended-twitter-accounts (last visited Oct. 23, 2021).

146. *My Account is Locked,* SNAPCHAT SUPPORT, https://support.snapchat.com/en-US/a/locked (last visited Oct. 23, 2012).

3. *Suspend Posting Rights*: Leave the account online but suspend the accountholder's ability to upload new content. Functionally, this can turn a "read-write" account into a "read-only" account. Temporarily limited posting rights can help "cool off" users.[147]

Examples:

- Reddit suspends violative users from posting, voting, commenting, and sending private messages.[148]
- Twitter limits some violative accounts so they cannot tweet, retweet, or like other posts.[149]
- Facebook restricts abusive users from using Facebook Live for set periods of time.[150]
- YouTube imposes a seven-day freeze on new uploads (plus other editing restrictions) for accounts that get a "strike" (i.e., a second warning).[151]

4. *Remove Credibility Badges*: Remove any service-provided badges or flair that enhance user credibility.

Example: Twitter removes its "blue check," which indicates that Twitter has verified the accountholder's identity, for "severe or repeated violation of the Twitter Rules."[152]

---

147. KRAUT & RESNICK, *supra* note 118, at 137. In LambdaMOO, one rule-breaker was given a "time out." Mnookin, *supra* note 35.

148. Catherine Shu, *Reddit Replaces Its Confusing Shadowban System with Account Suspensions,* TECHCRUNCH (Nov. 11, 2015, 11:11 PM), https://techcrunch.com/2015/11/11/reddit-account-suspensions.

149. *Help with Locked or Limited Account,* TWITTER HELP CTR., https://help.twitter.com/en/managing-your-account/locked-and-limited-accounts (last visited Oct. 23, 2021).

150. Guy Rosen, *Protecting Facebook Live from Abuse and Investing in Manipulated Media Research,* FACEBOOK NEWSROOM (May 14, 2019), https://about.fb.com/news/2019/05/protecting-live-from-abuse ("[A]nyone who violates our most serious policies will be restricted from using Live for set periods of time – for example 30 days – starting on their first offense.").

151. *Community Guidelines Strike Basics*, YOUTUBE HELP, https://support.google.com/youtube/answer/2802032?hl=en (last visited Oct. 23, 2021).

152. *Verification FAQ*, TWITTER HELP CTR., https://help.twitter.com/en/managing-your-account/twitter-verified-accounts (last visited Oct. 23, 2021). The stated reasons are: "Impersonation or intentionally misleading people on Twitter by changing your display name or bio; Violations that result in immediate account suspension; [or] Repeat violations in Tweets, including but not limited to: hateful conduct policy, abusive behavior, glorification of violence policy, civic integrity policy, private information policy, or platform manipulation and spam policy." *Id. See generally* Thomas Kadri, *Speech vs. Speakers*, SLATE (Jan. 18, 2018, 12:56 PM), https://slate.com/technology/2018/01/twitters-new-rules-blur-the-line-between-extremists-speakers-and-their-speech.html (explaining why de-verification might be counterproductive).

Twitter has since updated its "blue-check" policy. *Our Plans to Relaunch Verification and What's Next*, TWITTER BLOG (Dec. 17, 2020), https://blog.twitter.com/en_us/topics/company/2020/our-plans-to-relaunch-verification-and-whats-next.html.

5. *Reduce Service Levels*: The account can temporarily or permanently have less functionality or a lower quality of service. For example, the Internet service can limit the number of times an account's content can be read /viewed.

Examples:

- As discussed in Part II(A)(7), Internet access providers can "throttle" accounts as part of a graduated response.
- League of Legends places some users into low priority queues, lengthening the time it takes to join a new game.[153]
- LambdaMOO reduced violative players' storage space.[154]

6. *Shaming*: A service can publicly call attention to an accountholder's bad behavior.[155] This is similar to counterspeech for specific content items in the sense that it alerts readers of possible problems.

Examples:

- Yelp's "Consumer Alerts" program places warning badges on the pages of businesses that Yelp believes have tried to manipulate ratings or reviews.[156] Activities that can prompt warning badges include:
    - purchasing reviews or incentivizing people to write reviews;
    - writing reviews from the same IP address;
    - deceptive behavior;
    - media-fueled reviews; or
    - threatening reviewers with legal action.[157]
- The consumer review website Epinions issued "tickets" on the profile pages of violative accounts to signal the violation to the community.[158]

---

153. Laserface, *LeaverBuster FAQ,* RIOT GAMES (Apr. 14, 2011), https://support-leagueoflegends.riotgames.com/hc/en-us/articles/201752714.

154. Mnookin, *supra* note 35.

155. Schoenebeck et al, *supra* note 31, at 1295; Kate Klonick, *Re-Shaming the Debate: Social Norms, Shame, and Regulation in an Internet Age*, 75 MD. L. REV. 1029 (2016).

There has been extensive discussion of the pros and cons of shaming remedies in the criminal context. *See*, *e.g.*, *Incarceration Alternatives*, *supra* note 38, at 1957–67; Stephen P. Garvey, *Can Shaming Punishments Educate?*, 65 U. CHI. L. REV. 733 (1998); Kahan, *supra* note 41, 631–52; Toni M. Massaro, *Shame, Culture, and American Criminal Law*, 89 MICH. L. REV. 1880 (1991).

156. Vince Sollitto, *Yelp's Consumer Protection Initiative: Empowering Our Users*, YELP OFF. BLOG, (Aug. 1, 2018), https://blog.yelp.com/2018/08/yelps-consumer-protection-initiative-empowering-our-users.

157. *Id.* For more on Yelp's Consumer Alert program, see Noorie Malik, *How Yelp's Consumer Alerts Protected People from Misinformation in 2019*, YELP OFF. BLOG (Mar. 10, 2020), https://blog.yelp.com/2020/03/yelp-2019-consumer-alert-report.

- RuneScape showed players' discipline status in publicly visible "offence pillars."[159]
- Some virtual worlds turned violative players into virtual toads to humiliate them (called "toading").[160]

C. *Visibility Restrictions*

Internet services can downgrade the visibility of some or all of a user's content. The account and associated content remain available, but they may get less exposure.[161] This subpart describes eleven visibility restriction actions.

1. *Shadowban*: A shadowban keeps a user's account active, but only the accountholder can see the content.[162] Functionally, a shadowban resembles an account suspension, but: (1) a shadowbanned user can still access, edit, and download the content, and (2) users may not know they have been shadowbanned.[163] However, the term "shadowban" is used to describe other remedies, which has created substantial semantic confusion.[164]

---

158. *See Abuse and Site Rules*, EPIFAQ, http://epifaq.pbworks.com/w/page/11116355/Abuse_and_Site_Rules#ticketsdo (Jan. 4, 2007).

159. *Account Status*, RUNESCAPE WIKI, https://runescape.fandom.com/wiki/Account_Status (last visited Oct. 23, 2021).

160. Mnookin, *supra* note 35, at n.44.

161. Services can quantify an accountholder's reputation and reduce the visibility of low-reputation accountholders. "Karma" sometimes describes quantified reputations, and rule violations can be incorporated into a karma score to influence future visibility. F. RANDALL FARMER & BRYCE GLASS, BUILDING WEB REPUTATION SYSTEMS 72–73 (2010).

162. *E.g.*, G.F., *What is "Shadowbanning"?*, ECONOMIST (Aug. 1, 2018), https://www.economist.com/the-economist-explains/2018/08/01/what-is-shadowbanning ("Shadowbanned users are not told that they have been affected. They can continue to post messages, add new followers and comment on or reply to other posts. But their messages may not appear in the feed, their replies may be suppressed and they may not show up in searches for their usernames."); DeLima v. Google, Inc., No. 1:19-cv-978-JL, 2021 WL 294560, at n.13 (D.N.H. Jan. 28, 2021) ("Shadow banning is the act of blocking or partially blocking a user or their content from an online audience in a manner that is not readily apparent to the user. The user believes they are posting content normally, when in reality other people cannot see the posted content.").

For a history of shadowbanning and related remedies, see Samantha Cole, *Where Did the Concept of 'Shadow Banning' Come From?*, MOTHERBOARD: TECH BY VICE (July 31, 2018, 9:00 AM), https://www.vice.com/en_us/article/a3q744/where-did-shadow-banning-come-from-trump-republicans-shadowbanned.

163. *See* Nicolas P. Suzor et al., *What Do We Mean When We Talk About Transparency? Towards Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N. 1526, 1531 (2019).

164. *E.g.*, FLA. STAT. ANN. § 501.2041(1)(f) (West 2021) (defining "shadow ban" as "action by a social media platform, through any means, whether the action is determined by a natural person or an algorithm, to limit or eliminate the exposure of a user or content or material posted by a user to other users of the social media platform. This term includes acts of shadow banning by a social media platform which are not readily apparent to a user."); Chanté Joseph, *Instagram's Murky 'Shadow Bans' Just Serve to Censor Marginalised Communities*, GUARDIAN (Nov. 8, 2019, 11:03 AM), https://www.theguardian.com

Examples:

- A number of services have allegedly used shadowbanning, including Craigslist[165] and Reddit.[166]
- Allegedly at the Chinese government's request, TikTok displays some items only to the posting user.[167]
- Diaspora* hides the content of offenders from all users except the author and moderators.[168] Chatrooms can let a user see their chat messages but hide the messages from everyone else.[169] Similarly, gaming websites may "mute" abusive users and repeat offenders.[170]
- When content violates Blogger's policies, Blogger may unpublish the content so that it is visible only to its author.[171]

2. *Remove From External Search Index*: A service can place a "noindex" tag on a page so that the page does not appear in external search indexes

---

/commentisfree/2019/nov/08/instagram-shadow-bans-marginalised-communities-queer-plus-sized-bodies-sexually-suggestive (describing "shadowbans" as blocking Instagram hashtag searches and removal from Explore pages, both remedies discussed below).

Using a narrower definition of shadowbanning, President Trump accused Twitter of shadowbanning conservative politicians. Liam Stack, *What Is a 'Shadow Ban,' and Is Twitter Doing It to Republican Accounts?*, N.Y. TIMES (July 26, 2018), https://www.nytimes.com/2018/07/26/us/politics/twitter-shadowbanning.html.

165. On Craigslist, this remedy was called "ghosting." *E.g.*, Owais211, *What is Craigslist Ghosting and How to Fix it in 2018*, POSTING BROS (Jan. 12, 2018), https://clflaggingexpert.net/how-to-fix-craigslist-ghosting-issue.

166. *See* u/cojoco, *An Unofficial Guide on How to Avoid Being Shadowbanned*, REDDIT (Feb. 6, 2014), http://www.reddit.com/r/ShadowBan/comments/1x92jy/an_unofficial_guide_on_how_to_avoid_being; u/krispykrackers, *On Shadowbans.*, REDDIT (July 28, 2015), https://www.reddit.com/r/self/comments/3ey0fv/on_shadowbans ("A shadowban is the tool we currently use to ban people when they are caught breaking a rule. It causes their submitted content and user profile page to be visible only to themselves while logged in. Moderators can see their comments within their subreddit (since they can see 'removed' comments in the subreddit they moderate), but no other users can see their content, and nobody else can see their userpage."). Reddit abandoned shadowbanning in 2015. Shu, *supra* note 148.

167. Alex Hern, *Revealed: How TikTok Censors Videos That Do Not Please Beijing*, GUARDIAN (Sept. 25, 2019, 12:00 AM), https://www.theguardian.com/technology/2019/sep/25/revealed-how-tiktok-censors-videos-that-do-not-please-beijing.

168. WAGNER ET AL., *supra* note 105, at 17.

169. KRAUT & RESNICK, *supra* note 118, at 137 ("[I]n a chat room, the gagged person may see an echo of everything he or she types, but his or her comments may not be displayed to others in the room. The gagged person may think that everyone is just ignoring her.").

170. *E.g.*, Chipteck, *Chat Restrictions,* RIOT GAMES (Mar. 18, 2013, 4:40 PM), https://support.riotgames.com/hc/en-us/articles/201752984-Chat-Restrictions; u/Rocket_Sciencetist, *Muting Policy,* REDDIT (June 28, 2017), https://www.reddit.com/r/TagPro/comments/6k6k3o/muting_policy; *see also Mute,* FANDOM OLD SCHOOL RUNESCAPE WIKI, https://oldschoolrunescape.fandom.com/wiki/Mute (last visited Oct. 23, 2021).

171. *Blogger Content Policy*, BLOGGER, https://www.blogger.com/content.g?hl=en (last visited Oct. 23, 2021).

like Google, even though it remains fully accessible on the service.[172] The E.U.'s "Right to Be Forgotten"[173] provides an analogous remedy, though RTBF deindexing requests are submitted to search engines instead of the Internet services publishing the violative content.

Examples (from non-remedial contexts):

- Newspaper websites can no-index archival stories so that they do not appear in the search results for people named in the story.[174]
- Court websites can no-index court filings to make the filings available to the public but not visible to search engine searches on the referenced people's names.[175]

3. *Nofollow Authors' Links*: A service can place a "nofollow" tag on outlinks posted by users.[176] The "nofollow" tag tells Google and other search engines not to credit the link in their ranking algorithms.[177] This discourages users from posting links solely for search engine credit.

Example (from non-remedial context): Wikipedia puts nofollow tags on its outlinks to discourage the addition of links to its pages designed to generate marketing benefits, not to help readers.[178]

4. *Remove from Internal Search Index*: A service can remove content from its internal search index.

Examples:

- Reddit removes quarantined subreddits from its internal search.[179]

---

172. *Block Search Indexing with 'Noindex'*, GOOGLE SEARCH CENT., https://developers.google.com/search/docs/advanced/crawling/block-indexing (last updated Nov. 22, 2021).

173. Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, ECLI:EU:C:2014:317, ¶¶ 89–99 (May 13, 2014); *see Requests to Delist Content Under European Privacy La*w, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/eu-privacy/overview?hl=en (last visited Oct. 2, 2021).

174. *See* Zoe Greenberg, *Boston Globe Launches 'Fresh Start' Initiative: People Can Apply to Have Past Coverage About Them Reviewed*, MSN (Jan. 22, 2021), https://www.msn.com/en-us/news/us/boston-globe-launches-fresh-start-initiative-people-can-apply-to-have-archive-stories-about-them-reviewed/ar-BB1cYPrp.

175. *See*, *e.g.*, Eugene Volokh, *Can You Get a Court to Take an Opinion That Mentions You Off Its Google-Searchable Website?*, WASH. POST (May 3, 2017), https://www.washingtonpost.com/news/volokh-conspiracy/wp/2017/05/03/can-you-get-a-court-to-take-an-opinion-that-mentions-you-off-its-google-searchable-website.

176. KRAUT & RESNICK, *supra* note 118, at 154–55.

177. *Id.*; *Qualify Your Outbound Links to Google*, GOOGLE SEARCH CENT., https://support.google.com/webmasters/answer/96569?hl=en (last updated Aug. 26, 2021).

178. Bobbie Johnson, *Wikipedia Adopts 'Nofollow'*, GUARDIAN (Jan. 22, 2007, 7:19 AM), https://www.theguardian.com/technology/blog/2007/jan/22/wikipediaadopt. *See generally* Eric Goldman, *Wikipedia's Labor Squeeze and Its Consequences*, 8 J. TELECOMM. & HIGH TECH. L. 157, 163 (2010) (discussing Wikipedia's efforts to suppress spam).

- Instagram removes false posts from its hashtag search.[180]

5. *Downgrade Internal Search Visibility*: Instead of removing content entirely from its internal index, a service can downgrade content's visibility on the internal search results page.

Example: Facebook downgrades pages and content that are sensational, spammy, or misleading.[181]

6. *No Auto-Suggest*: A service's internal search engine can remove content from its "auto-suggest" search feature.

Examples:

- Google Search blocks autocompletes for a variety of terms, such as words allegedly associated with copyright infringement.[182]
- In 2018, Twitter removed some accounts from its auto-suggest feature.[183]

7. *No/Reduced Internal Promotion*: Many services do internal cross-promotions, including recommendations. To reduce its exposure, a service can remove or downgrade content from one or more of these internal promotions.

Examples:

- YouTube does not recommend borderline videos.[184]

---

179. *Quarantined Subreddits, supra* note 133.

180. *Combatting Misinformation on Instagram*, *supra* note 138.

Instagram allegedly penalizes users by removing their posts from hashtag searches, which is sometimes called Instagram "shadowbanning." Taylor Lorenz, *Instagram's "Shadowban," Explained: How to Tell if Instagram Is Secretly Blacklisting Your Posts*, MIC (June 7, 2017), https://www.mic.com/articles/178987/instagrams-shadowban-explained-how-to-tell-if-instagram-is-secretly-blacklisting-your-posts.

181. In 2019, Facebook updated its ranking algorithms to "reduce (1) posts with exaggerated or sensational health claims and (2) posts attempting to sell products or services based on health-related claims." Travis Yeh, *Addressing Sensational Health Claims,* FACEBOOK NEWSROOM (July 2, 2019), https://about.fb.com/news/2019/07/addressing-sensational-health-claims; *People, Publishers, the Community,* FACEBOOK NEWSROOM (Apr. 10, 2019), https://about.fb.com/news/2019/04/people-publishers-the-community.

182. Barry Schwartz, *Google Expanding Types of Predictions They Remove from Autocomplete*, SEARCH ENGINE LAND (Apr. 20, 2018, 12:47 PM), https://searchengineland.com/google-expanding-types-of-predictions-they-remove-from-autocomplete-296576.

183. Twitter claimed the removals were due to a bug. Vijaya Gadde & Kayvon Beykpour, *Setting the Record Straight on Shadow Banning*, TWITTER BLOG (July 26, 2018), https://blog.twitter.com/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html.

184. *Continuing Our Work to Improve Recommendations on YouTube*, YOUTUBE OFF. BLOG (Jan. 25, 2019), https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html ("[W]e'll begin reducing recommendations of borderline content and content that could misinform users in harmful ways. . . this will only affect recommendations of what videos to watch, not whether a video is available on YouTube."). This also applies to

- Twitter does not algorithmically recommend tweets that violate Twitter's policies but remain posted due to the public interest.[185]
- Facebook reduces News Feed visibility of pages that display false content.[186]

8. *No/Reduced Navigation Links*: Many services provide lists of links such as "most popular" or "newly available" items. To reduce its exposure, a service can remove or downrank content from one or more of these lists.

Examples:

- Quarantined subreddits do not appear in "non-subscription-based feeds."[187]
- Instagram removes false posts from its "Explore" pages.[188]

9. *Reduced Virality*[189]: Social media services can limit the ability of users to share content by adding friction to the sharing process or blocking inter-user sharing altogether.

Example: Twitter has attempted to reduce virality of some violative tweets by restricting the ability of other users to retweet, like, or share the tweets.[190]

10. *Age-Gate*: A service may restrict minors' access to content.

Example: YouTube users can opt-into a "restricted mode" (also called "safe mode"), which blocks the visibility of "mature" videos.[191]

11. *Display Content Only to Logged-In Readers*: A service can show content only to registered readers. This hides the content from unregistered users, such as first-time visitors and visitors referred by search engines. Often, a site's registered users are a small fraction of its total audience, so hiding the content from unregistered users can significantly reduce the audience for that content.

---

inflammatory religious or supremacist content. Walker, *supra* note 134 (applying recommendation reductions to inflammatory religious or supremacist content).

185. *About Public-Interest Exceptions on Twitter*, *supra* note 132.

186. *Fact Checking on Facebook,* FACEBOOK FOR BUS., https://www.facebook.com/help/publisher/182222309230722 (last visited Oct. 24, 2021).

187. *Quarantined Subreddits, supra* note 133.

188. *Combatting Misinformation on Instagram*, *supra* note 138.

189. Other remedies discussed in this subpart also can help decelerate virality.

190. Vijaya Gadde & Kayvon Beykpour, *Additional Steps We're Taking Ahead of the 2020 US Election*, TWITTER BLOG, https://blog.twitter.com/en_us/topics/company/2020/2020-election-changes.html (Nov. 2, 2020); *About Public-Interest Exceptions on Twitter*, *supra* note 132.

191. *Turn Restricted Mode On or Off*, YOUTUBE HELP, https://support.google.com/youtube/answer/174084?co=GENIE.Platform%3DDesktop&hl=en (last visited Oct. 2, 2021); *see also* Prager Univ. v. Google LLC, 951 F.3d 991, 996 (9th Cir. 2020) (discussing YouTube's restricted mode).

Example: Epinions let its registered users rate other users' reviews as "very helpful," "helpful," "somewhat helpful," or "not helpful."[192] Reviews with a net rating of "somewhat helpful" or "not helpful" were shown only to logged-in readers.[193] Thus, for a review to reach the site's full audience, it needed a net user rating of "very helpful" or "helpful."

D. *Monetary*

Where services pay authors for content or hold their users' money, the following four additional remedies become viable.

1. *Forfeit Accrued Earnings*: A service can withhold any accrued earnings.

Example: Google withholds accrued earnings for publishers who violate its AdSense rules.[194]

2. *Terminate Future Earnings (By Item or Account)*: A service can terminate future payments, sometimes called "demonetization."[195]

Examples:

- YouTube may terminate payments for individual videos or entire channels.[196]
- As part of its Content ID program, YouTube allows copyright owners to claim the revenues from an allegedly infringing work, effectively assigning future earnings to the copyright owner.[197]
- Quarantined subreddits do not earn revenue.[198]

3. *Suspend Future Earnings (By Item or Account)*: Instead of permanently terminating the ability to earn, a service can temporarily suspend that ability.

---

192. *Reviews*, EPIFAQ, http://epifaq.pbworks.com/w/page/11116374/Reviews#nomoney (last visited Oct. 2, 2021).

193. *Ratings and the Web of Trust*, EPIFAQ, http://epifaq.pbworks.com/w/page/11116373/Ratings_and_Web_of_Trust#communitystandards (last visited Oct. 2, 2021). Slashdot uses a similar scoring system to hide lowly-rated user comments from public view. *See* WAGNER ET AL., *supra* note 105, at 20.

194. *Why Your AdSense Account Has a Payment Hold*, GOOGLE ADSENSE HELP, https://support.google.com/adsense/answer/1714364?hl=en (last visited Oct. 24, 2021).

195. *E.g.*, Julia Alexander, *YouTube Looks to Demonetization as Punishment for Major Creators, But It Doesn't Work*, VERGE (June 25, 2019, 12:27 PM), https://www.theverge.com/2019/6/25/18744246/youtube-demonetization-steven-crowder-patreon-advertising-merch.

196. Benjamin Goggin & Kat Tenbarge, *'Like You've Been Fired from Your Job': YouTubers Have Lost Thousands of Dollars After Their Channels Were Mistakenly Demonetized for Months*, BUS. INSIDER (Aug. 24, 2019, 9:31 AM), https://www.businessinsider.com/youtubers-entire-channels-can-get-mistakenly-demonetized-for-months-2019-8. YouTube also demonetizes extremist religious or supremacist content. Walker, *supra* note 134.

197. *What is a Content ID Claim?*, YOUTUBE HELP, https://support.google.com/youtube/answer/6013276?hl=en (last visited Oct. 2, 2021).

198. *Quarantined Subreddits, supra* note 133.

Example: Epinions paid authors for reviews that were rated "helpful" or "very helpful" by other reviewers.[199] Reviews rated "somewhat helpful" or "not helpful" were unpaid so long as they held that status,[200] but they could resume earnings if their ratings improved.[201]

4. *Fine Author/Impose Liquidated Damages*: A service can financially penalize violators, either by (1) taking some or all of the user's money in the service's possession (this may overlap with the forfeit remedy), or (2) by imposing a "fine" ("liquidated damages" if specified in the TOS).[202]

Examples:

- MySpace imposed liquidated damages on users who spammed.[203]
- Ticketmaster imposed liquidated damages on unauthorized bot purchases of tickets.[204]

E. *Other*

Seven other remedies that do not fit into the prior categories.

1. *Educate Users*: A service can treat rule violations as opportunities to teach the user about the service's rules and norms.[205]

Examples:

- The Copyright Alert System's remedies included user education.[206]
- League of Legends uses "reform cards" and abuse reports to give timely feedback to violative users.[207]

---

199. *Reviews*, *supra* note 192.

200. *Id.*

201. This payment approach reduced the submission of low-quality reviews because they were not profitable. *See infra* Part IV(B)(3).

202. KRAUT & RESNICK, *supra* note 118, at 161–62; *see also* Grimmelmann, *supra* note 12, at 68.

203. MySpace, Inc. v. Globe.com, Inc., No. CV 06-3391-RGK (JCx), 2007 U.S. Dist. LEXIS 44143, at *28–32 (C.D. Cal. Feb. 27, 2007).

204. TicketMaster LLC v. Prestige Ent., Inc., 306 F. Supp. 3d 1164, 1176–77 (C.D. Cal. 2018).

205. Here is an example, though it involved an effort independent from the Internet service. A campaign called "We Counter Hate" used AI to identify potentially hateful tweets. A human then responded to the Twitter user: "This hate tweet is now being countered. Think twice before retweeting. For every retweet, a donation will be committed to a non-profit fighting for equality, inclusion, and diversity." Cosette Jarrett, *AI Could Make Trolls Think Twice Before Retweeting Offensive Content*, VENTUREBEAT (Feb. 4, 2018, 10:19 PM), https://venturebeat.com/2018/02/04/ai-could-make-trolls-think-twice-before-retweeting-offensive-content. For more examples, *see* WAGNER ET AL., *supra* note 105, at 16–23.

206. *See supra* Part II(A)(7).

2. *Assign Strikes/Warnings*: A service can warn users after rule violations and track those warnings using strikes.

Examples:

- The graduated response schemes used strikes and warnings.[208]
- All services seeking to qualify for the DMCA online safe harbor assign strikes.[209]
- In YouTube's strike system, the first violation typically gets a warning. The first "strike" occurs on the second violation.[210]

3. *Outing/Unmasking*: A service can reveal a pseudonymous user's identity, which can lead to shaming (discussed above) or other judicial or extra-judicial consequences.[211]

Examples:

- 17 U.S.C. § 512(h) (part of the DMCA) provides an expedited "outing" procedure for alleged copyright infringers. After sending a takedown notice, the copyright owner can obtain a subpoena to unmask the alleged infringer. The court clerk must issue the subpoena without further judicial review.[212]
- The U.K. Defamation Act requires services to provide the contact information of users to complainants to avoid defamation liability.[213]

4. *Report to Law Enforcement*: A service may report a violation to law enforcement for possible prosecution. This remedy likely complements other remedies, including content removal and account termination.

Examples:

- Internet services must notify the National Center for Missing and Exploited Children (NCMEC) about any child sexual abuse material (CSAM) they discover on their networks.[214]
- Australia requires services to notify law enforcement if they learn about livestreaming of certain crimes.[215]

---

207. Brendan Maher, *Good Gaming*, 531 NATURE 568, 570 (Mar. 30, 2016), https://www.nature.com/articles/531568a.

208. *See supra* Part II(A)(7).

209. *See supra* Part II(A)(1).

210. *Community Guidelines Strike Basics, supra* note 151*; see also* YouTube Report, *supra* note 22, at 16.

211. *See infra* Part IV(A)(7) (revisiting the implications of extra-judicial consequences from outing).

212. 17 U.S.C. § 512(h).

213. *See supra* Part II(A)(2).

214. 18 U.S.C. § 2258A.

215. *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) (Austl.).

5. *Put User/Content on Industry-Wide Blocklist*: A service can share information with other services about violations. This could lead to industry-wide "blocklists" for users or specific content items.

Examples:

- "Global Internet Forum for Countering Terrorists" (GIFCT) is an industry-wide blocklist for photos and videos that a participating service has identified as terrorist content.[216]
- Uber and Lyft created the "Industry Sharing Safety Program," a blocklist of drivers accused of sexual or physical abuse.[217]

6. *"Community Service"*: A service can require a user to perform some service to the community to regain good standing.

Example: Community service was a remedy in LambdaMOO.[218]

7. *"Restorative Justice"/Apology*: A violation often affects community members, not just the service. To redress the harm caused to the community, a violating user could apologize to affected users[219] or participate in a more robust restorative justice process, such as a community discussion about how the violation affected the community.

Examples:

- r/Christianity subreddit moderators experimented with restorative justice. They paired abusive users with mediators in private chatrooms to discuss why their content was problematic.[220]
- Voluntarily made apology videos have become a genre among YouTubers.[221]

F. *Combining Remedies*

The prior subpart discussed each remedy in isolation, but remedies can be combined to increase their efficacy. For example, for Quarantined Communities, Reddit imposes multiple remedies simultaneously:

---

216. GLOBAL INTERNET FORUM TO COUNTER TERRORISM, https://gifct.org (last visited Oct. 2, 2021).

217. I. Bonifacic, *Uber and Lyft Create a Shared Database of Drivers Banned for Assault*, ENGADGET (Mar. 11, 2021), https://www.engadget.com/uber-lyft-industry-sharing-safety-program-204433080.html.

218. Mnookin, *supra* note 35.

219. In a survey, some (but not all) victim communities supported apologies as remedies for online harms. Schoenebeck et al., *supra* note 31, at 1289 tbl.4, 1293–95.

220. Charlie Warzel, *Could Restorative Justice Fix the Internet?,* N.Y. TIMES (Aug. 20, 2019), https://www.nytimes.com/2019/08/20/opinion/internet-harassment-restorative-justice.html.

221. Bettina Makalintal, *How YouTubers Turned the Apology Video Into a Genre,* VICE (June 18, 2019, 1:53 PM), https://www.vice.com/en_us/article/ywykzb/how-youtubers-james-charles-jaclyn-hill-pewdiepie-turned-the-apology-video-into-a-genre.

> Quarantined communities will display a warning that requires users to explicitly opt-in to viewing the content. They generate no revenue, do not appear in non-subscription-based feeds (eg Popular), and are not included in search or recommendations. Reddit may also enforce a number of additional product restrictions that exist currently or as they may develop in the future (eg removing custom styling tools).[222]

When developing a remedial strategy, services should evaluate remedies both in isolation and in combination. The efficacy of remedy combinations will likely vary by community and violation type, and the best answers will come only from experimentation and empirical data. Still, it seems inevitable that sometimes remedy combinations will work better than individual remedies in isolation, much like how chemotherapy can be more effective when drugs are used in combination than any single drug can achieve on its own.

IV. PRIORITIZING REMEDY OPTIONS

Part III described many remedy options. This Part tackles the natural follow-up question: how should regulators and Internet services navigate these options? In other words, if we move away from the binary thinking about remedies for rule violations, what practical, normative, or philosophical principles should guide the choices among the universe of remedy options?

There is no single ideal solution for the design and implementation of remedial schemes.[223] First, each solution reflects normative views that are not universally shared, so ideological conflicts are unavoidable.[224] Second, competing values may contradict each other, so choosing between those values will necessitate unwanted tradeoffs.[225] Third, because services' communities differ from each other, remedies that work in one community may not work elsewhere. Thus, rather than articulate a single "solution" to

---

222. *Quarantined Subreddits*, *supra* note 133.

223. *See* Schoenebeck et al., *supra* note 31 (presenting survey results showing that users had diverse feelings about the appropriateness of different remedies); *see also* Sarita Schoenebeck et al., *Beyond Borders: Women's Perspectives on Harm and Justice after Online Harassment* (Apr. 2021) (unpublished manuscript) (on file with the Michigan Technology Law Review) (finding that women's preferred remedies varied by geographic region and harm type).

224. Grimmelmann, *supra* note 12, at 70. This resembles the debates between criminal law scholars over the merits of deterrence versus retribution and how different criminal remedies might advance one norm better (or at the expense of) the other norm.

225. Douek, *supra* note 33, at 769 ("Online speech governance is a wicked problem with unenviable and perhaps impossible trade-offs."); *cf.* John M. Golden, *Principles for Patent Remedies*, 88 TEX. L. REV. 505, 552 (2010) (discussing how principles for remedies design "might point in different directions").

the unsolvable remedy prioritization challenge, this Part enumerates the considerations that should guide the decision-making.

A. *Factors to Consider*

This subpart explores some factors that regulators and Internet services can evaluate when setting policy about remedy options. The factors cannot be rank-ordered because no single factor is "best" in the abstract. However, an option could be "best" for a particular service or regulator in a particular circumstance. In practice, many of these factors will be simultaneously in play with each remedy option, and often the factors will need to be balanced or traded-off against each other.

Some factors for choosing among remedies include:

- severity of the rule violation;
- confidence that a rule violation actually occurred;
- scalability and consistency;
- the community's ability to self-correct;
- how the remedies impact others;
- retaining user engagement while curbing violations and recidivism;
- parallel sanctions.[226]

This subpart examines each factor in more detail:

1. *Severity of the Rule Violation*. The remedy should be proportionate to the rule violation.[227] More severe violations should trigger more significant remedies.[228]

In practice, only a narrow band of activity may be subject to discretion. If no rule violation has occurred, then remedies are not necessary at all.[229]

---

226. This list is meant to be illustrative, not exhaustive. For example, Google/YouTube says they "value openness and accessibility," "respect user choice," and "build for everyone." YouTube Report, *supra* note 22, at 3. Elsewhere, it notes the importance of worker wellness. *Id.* at 16–18.

227. Land & Hamilton, *supra* note 108, at 148 ("Proportional responses are required both by international human rights law as well as the law of remedies in international law more generally"). Marique & Marique summarize the principles:

> [T]he proportionality test implies that decision-makers should only follow a course of action if: 1) their objective is legitimate; 2) their means is necessary to achieve the objective; 3) no means would entail a lighter encroachment of the right at stake; 4) the means is proportional (*sensu stricto*) to the objective to be achieved.

Marique & Marique, *supra* note 28, at 9.

228. *Oversight Bd. Decision*, *supra* note 144 (stating that Facebook must consider "the gravity of the violation and the prospect of future harm" when determining remedies); KRAUT & RESNICK, *supra* note 118, at 162–63; Land & Hamilton, *supra* note 108, at 148 ("Any remedy chosen must be proportional to the gravity and harm of the violation"); *see also* YouTube Report, *supra* note 22, at 20 (referring to some matters as "Your Money or Your Life (YMYL)").

On the other side, some rule violations are severe enough to justify automatic removal. Child sexual abuse material (CSAM) is a paradigmatic example of content that should always be removed as quickly as possible. More generally, we might start with a rebuttable presumption that violations of government-made law are more severe than violations of house rules, though there could be exceptions in both directions.

Severity can be measured as a spectrum ranging from 0 to 100. 0 represents no rule violation at all; 100 is the worst possible violation. Services will likely pick a threshold number (x), something less than 100, where removal automatically applies. That leaves the range from 1 to x as the relevant range for non-removal remedies. Within that range, the service should scale remedies proportionately, i.e., the closer to x, the more severe the remedy (but less than complete removal).

2. *Confidence That a Rule Violation Actually Occurred*. It will not always be clear that a rule violation occurred. CSAM is comparatively unique in this regard; violations usually can be confirmed by reference to the content item.[230]

In contrast, in many circumstances, a rule violation cannot be definitively determined.[231] For example, take a situation where the applicable rules restrict defamatory content. For a service to determine if a user-supplied statement is defamatory, it will need to decide if the statement is true or false. However, the information needed to decide that question often will not be available to the service. As a result, a service deciding if a user-supplied statement is defamatory will have to make a (hopefully educated) guess.

In practice, services routinely impose remedies for rule violations when they are not 100% sure that a rule violation took place. For example, the DMCA's notice-and-takedown provision pushes services to remove user content based on unproven assertions that infringement took place, without conducting any independent research to validate the claims in the notice (and knowing that many claims are, in fact, false).[232]

---

229. If the activity is nevertheless anti-social or otherwise harms the community, a service might reevaluate its rules to restrict it.

230. This comparative ease of detection has contributed to the effectiveness of filters such as PhotoDNA. *See* Klonick, *supra* note 19, at 1636–37.

231. *See* Goldman & Miers, *supra* note 21, at 204–07.

232. *See*, *e.g.*, Jennifer M. Urban, Joe Karaganis & Brianna L. Scholfield, *Notice and Takedown in Everyday Practice*, 40–43 (U.C. Berkeley Pub. Law Rsch. Paper, Paper No. 2755628, 2017), https://ssrn.com/abstract=2755628. 17 U.S.C. § 512(f) sought to increase services' confidence that copyright owners were only sending notices in the cases of actual infringement by penalizing bogus takedown notices. However, due to drafting errors, Section 512(f) does not effectively discipline copyright owner misbehavior. *See* Eric Goldman, *How Have Section 512(f) Cases Fared Since 2017? (Spoiler: Not Well)*, TECH. & MKTG. L. BLOG (Apr. 6, 2019), https://blog.ericgoldman.org/archives/2019/04/how-have-section-512f-cases-fared-since-2017-spoiler-not-well.htm.

Services could impose remedies only after definitive proof that a violation occurred. Such proof could come from a third-party adjudicator, such as a court, or through independent investigation by the service until it has reached an irrefutable conclusion. Indeed, Congress has proposed to mandate increased investigatory obligations by services[233] nominally in support of due process values. However, requiring services to confirm rule violations before imposing remedies has its own downsides. Services must either incur potentially high investigatory costs (in tension with the scalability principle discussed below),[234] or services will not take action because it is impossible to confirm rule violations.

Non-removal remedies provide possible workarounds to this conundrum.[235] Where a service suspects, but cannot prove, that a rule violation took place, the service might deploy less severe remedies.[236] This has several benefits. First, it reduces the risks of "false positive" removals. Second, the service could use disclosure-focused remedies—such as fact-checks or interstitial warnings—to signal its uncertainty. Non-removal remedies preserve the opportunity for helpful counterspeech, such as corrective reader comments.[237] Third, non-removal remedies may allow additional facts to emerge, which could help the service make a more accurate decision later.[238] Fourth, implementing non-removal remedies in response to unproven allegations reduces the ability of malefactors to successfully game or weaponize the removal remedy to achieve illegitimate outcomes.[239]

Striking a balance between remedy imposition and confidence of a rule violation also arises in courts' imposition of preliminary and permanent

---

233. *E.g.*, Online Content Policy Modernization Act, S. 4632, 116th Cong. § 201(1)(A) (2020) (proposing to remove the legal protections in 47 U.S.C. § 230(c)(2)(A) unless the service has an "objectively reasonable belief" that any removed content or accounts fit into one of the specified content categories).

234. Goldman & Miers, *supra* note 21, at 205–06.

235. Douek, *supra* note 33.

236. *Id.* at 789–800 (advocating for content moderation systems built on the premise that errors are unavoidable); *cf.* Federico Picinali, *Do Theories of Punishment Necessarily Deliver a Binary System of Verdicts? An Exploratory Essay*, 12 CRIM. L. & PHIL. 555 (2018) (discussing how the criminal system could calibrate verdicts to standards lower than "beyond a reasonable doubt"); Mark Spottswood, *Continuous Burdens of Proof*, 21 NEV. L.J. 779, 829 (2021) (discussing how burdens of proof could be a continuum rather than binary options).

237. Douek, *supra* note 33, at 816.

238. For example, the Wikimedia community has repeatedly resolved copyright disputes through its own independent research without intervention by Wikimedia employees or the courts. *See Stories*, WIKIMEDIA FOUND. TRANSPARENCY REP., https://transparency-archive.wikimedia.org/stories.html (last visited Oct. 25, 2021).

239. Internet services' willingness to over-remove has been weaponized in the past. *See* V. Blue, *Why PayPal's Crackdown on ASMR Creators Should Worry You*, ENGADGET (Sept. 14, 2018), https://www.engadget.com/2018-09-14-paypal-ban-asmr-sound-art-therapy.html (malefactors got PayPal to ban legitimate content producers by sending bogus takedown notices).

injunctions.[240] A court may issue a preliminary injunction when the plaintiff is likely to succeed (or, sometimes, based on even lower confidence of success), even though the plaintiff has not yet proven that a legal violation took place.[241] At the same time, because the plaintiff's merits are not definitively resolved, a court may tailor any preliminary relief to reflect the balance of equities and the public interest.[242] In contrast, with a permanent injunction, the court knows that the defendant's legal violation has already been shown. The public interest is still relevant, but other factors emphasize the need for remediation.[243]

When a service has irrefutable proof, or a very high degree of confidence, that a rule violation has taken place, it is closer to a permanent injunction. However, when a service has less confidence about the rule violation's occurrence, the preliminary injunction analogy fits better.

3. *Scalability and Consistency*. Services usually aspire to scalable yet consistent content moderation processes,[244] including remedies.[245] Many services would prefer, in theory, to make individualized remedy determinations after taking account of all facts and circumstances.[246] In practice, this is not possible due to the high cost of individualized remedies and the risk of inconsistent outcomes. Inconsistency hurts the individuals who get the harsher remedies and undermines users' and regulators' confidence in the service's legitimacy.[247] Inconsistency can also stem from

240. Similarly, the proof standard for criminal conviction ("beyond a reasonable doubt") is higher than the standard for civil decisions (e.g., "preponderance of the evidence"), in part because criminal sanctions may be more consequential.

241. Winter v. Nat. Res. Def. Council, Inc., 555 U.S. 7, 20 (2008) ("A plaintiff seeking a preliminary injunction must establish that he is likely to succeed on the merits, that he is likely to suffer irreparable harm in the absence of preliminary relief, that the balance of equities tips in his favor, and that an injunction is in the public interest."); *see also* CHARLES ALAN WRIGHT, ARTHUR R. MILLER & MARY KAY KANE, FEDERAL PRACTICE AND PROCEDURE § 2948.3 (3d ed. 2020) (enumerating the "bewildering variety of formulations" courts consider for the plaintiff's burden of proof for preliminary relief).

242. *See* Winter v. Nat. Res. Def. Council, Inc., 555 U.S. at 24–33.

243. *See* eBay Inc. v. MercExchange, L.L.C., 547 U.S. 388 (2006) ("A plaintiff must demonstrate: (1) that it has suffered an irreparable injury; (2) that remedies available at law, such as monetary damages, are inadequate to compensate for that injury; (3) that, considering the balance of hardships between the plaintiff and defendant, a remedy in equity is warranted; and (4) that the public interest would not be disserved by a permanent injunction.").

244. Douek, *supra* note 33, at 791 ("Scale is major platforms' Prime Directive . . . .").

245. In jurisprudential parlance, "administrability" might be a synonym. *See* Golden, *supra* note 225, at 512.

246. Tarleton Gillespie referred to this as "artisanal" content moderation. GILLESPIE, *supra* note 19, at 77; *see also* ROBYN CAPLAN, DATA & SOC'Y, CONTENT OR CONTEXT MODERATION: ARTISANAL, COMMUNITY-RELIANT, AND INDUSTRIAL APPROACHES 17–19 (2018), https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf.

247. *See* Common, *supra* note 16, at 138–41 (discussing the virtues of consistency in content moderation).

unwanted biases[248] in the content moderation process, including discriminatory animus or effect based on intrinsic characteristics, something that most services try to avoid (and may be legally required to avoid[249]).

In modeling the tradeoffs between scalability and consistency, consider a hypothetical example where a service expects content reviewers to process potential rule violations once every minute on average. The service then presents the content reviewer with a menu of remedy options for an identified rule violation and reviewing this menu and choosing a customized remedy adds another ten seconds to the review. In this example, the remedy menu increases the workload over 15% for each identified violation—a potentially costly burden.[250] Furthermore, the additional remedial choices create more opportunities for content reviewers to reach inconsistent conclusions.

"Scalability" is not intrinsically a positive value.[251] However, because it addresses concerns about cost-effectiveness and consistency, scalability is critical to Internet services.

4. *The Community's Ability to Self-Correct.* In tight-knit communities where participants are repeat players, the community may be able to self-discipline rule violations.[252] If so, non-removal remedies might helpfully supplement the community's own responses. Less tight communities may not self-correct as easily, pushing the service to intervene more aggressively.[253] Then again, Wikipedia has built self-policing into its design, and its openness increases the odds of successful community self-correction.

5. *How the Remedies Impact Others*. A service's imposition of remedies can have substantial implications for others inside and outside the community.[254] The remedial scheme should reflect these considerations. The underlying rule may seek to benefit:

- a specific victim (e.g., an anti-defamation or anti-copyright infringement rule);
- the Internet service itself (e.g., a rule against consuming too many system resources);

248. The word "bias" is in quotes because all service editorial decisions are unavoidably "biased." *See generally* Eric Goldman, *Search Engine Bias and the Demise of Search Engine Utopianism*, 8 YALE J.L. & TECH. 188 (2006).

249. *E.g.*, Harrington v. Airbnb, Inc., 348 F. Supp. 3d 1085, 1089–92 (D. Ore. 2018).

250. The turnaround time to implement remedies may be another consideration. As the old maxim goes, "justice delayed is justice denied."

251. Indeed, it can be the source of considerable concern. *See* Common, *supra* note 16, at 135–38 (criticizing the "narrative of efficiency").

252. KRAUT & RESNICK, *supra* note 118, at 140; *see also* ROBERT C. ELLICKSON, ORDER WITHOUT LAW: HOW NEIGHBORS SETTLE DISPUTES 211–19 (Harvard Univ. Press 1991).

253. *E.g.*, Neil Weinstock Netanel, *Cyberspace Self-Governance: A Skeptical View from Liberal Democratic Theory*, 88 CALIF. L. REV. 395, 429–32 (2000).

254. *E.g.*, Perel, *supra* note 40, at 30–35.

- the service's community (e.g., pro-civility rules); or
- the public generally (e.g., a rule against election interference).

Services are well-positioned to decide what remedies will best balance the competing interests when the intended rule beneficiaries are themselves or their communities. In contrast, services are not well-positioned to understand the needs of specific victims or the public generally. In those situations, Internet services are in the impossible position of trying to balance interests that they may not understand and that may irreconcilably conflict with each other.[255] Yet, in those circumstances, the Internet services will inevitably prioritize their own interests and profits.

While Internet services may not care directly about the consequences of their remedial actions on specific victims or the public at large—especially when Section 230 negates their legal exposure[256]—Internet services cannot ignore these consequences either. In some circumstances, removal is the only tenable option to eliminate the harm. In others, non-removal remedies help balance the competing/conflicting interests, such as the author's free expression, while still benefitting external parties. However, ideally Internet services will design those remedies in consultation with the affected parties so that the services can better understand their needs.

6. *Retaining User Engagement While Curbing Violations and Recidivism*. Most Internet services prefer to rehabilitate users rather than banish them. Imposing remedies on a user runs the risk of driving the user away or suppressing their engagement, but the remedies also need to discourage recidivism.[257] Services might choose to impose remedies that balance rehabilitation and anti-recidivism with future engagement.

The visibility of imposed remedies has potentially significant implications. Publicly imposing remedies can enhance deterrence.[258] It can

255. The no-win nature of content moderation decisions motivated Facebook to create the Oversight Board. Klonick, *supra* note 19, at 2427–48. Essentially, Facebook outsources its no-win decisions to the Oversight Board—and lets the board take the heat from people unhappy with its moderation decisions. *See generally* Eric Goldman, *Top Myths About Content Moderation*, TECH. & MKTG. L. BLOG (Oct. 15, 2019), https://blog.ericgoldman.org/archives/2019/10/top-myths-about-content-moderation.htm (discussing the no-win nature of content moderation).

256. Goldman, *Section 230 Overview*, *supra* note 5, at 158–60.

257. *E.g.*, *Oversight Bd. Decision*, *supra* note 144, at 6 ("Suspension periods should be long enough to deter misconduct . . . ."); Jhaver et al., *supra* note 12, at 20–23; Land & Hamilton, *supra* note 108, at 149 ("[A] remedy includes the duty to take appropriate measures to prevent future violations . . . .").

258. KRAUT & RESNICK, *supra* note 118, at 143; Joseph Seering, Robert Kraut & Laura Dabbish, *Shaping Pro and Anti-Social Behavior on Twitch Through Moderation and Example-Setting*, *in* CSCW'17 PROCEEDINGS OF THE 2017 ACM CONFERENCE ON COMPUTER SUPPORTED COOPERATIVE WORK AND SOCIAL COMPUTING 111, 112 (2017) (discussing "behavioral imitation, where observing one type of behavior encourages observers to behave in the same way"). As Seering et al. explain, "[m]oderation can be viewed not only

also bolster community trust by demonstrating that the service takes rule enforcement seriously;[259] and it can signal virtue (i.e., the service is acting "tough") to placate regulators or advocacy groups. On the other hand, a publicly visible remedy might counterproductively raise awareness of rule-violating content,[260] and a less public remedy might have better odds of rehabilitating a violative user.[261]

With respect to rehabilitation and recidivism, Ayres & Braithwaite advocated for imposing discipline that becomes progressively more severe.[262] Braithwaite described a "pyramid of sanctions"[263] (this figure applies to selling medicines[264]):

---

as a reaction to specific events but also a method for preventing the spread of unwanted behavior and development of undesirable norms for what conduct is acceptable." *Id.* at 124.

259. Grimmelmann, *supra* note 12, at 65–66. Two examples of highly visible enforcement efforts designed to appeal to other users:

- EVE Online punished cheaters by placing their spaceships in a public space where other users could easily kill them off. Lee Yancy, *An Official EVE Online Event Let Players Publicly 'Execute' Cheaters*, KOTAKU (Aug. 30, 2018, 5:00 PM), https://www.kotaku.com.au/2018/08/an-official-eve-onlineevent-let-players-publicly-execute-cheaters. This public spectacle satisfies other players' desires for vengeance and reinforces EVE Online's anti-cheating stance.
- RuneScape lets other players vote how to destroy a bot player's avatar. *See* Tom Senior, *RuneScape Puts Botters on Trial in Botany Bay and Lets Players Decide Their Fate*, PC GAMER (Sept. 26, 2012), https://www.pcgamer.com /runescape-to-get-a-botmaster-general-and-put-botters-on-trial-in-botany-bay.

260. KRAUT & RESNICK, *supra* note 118, at 144–45. This is a variation of the Streisand Effect. *See* Guttenberg v. Emery, 26 F. Supp. 3d 88, 95 (D.D.C. 2014); Bill Mordan, *The Streisand Effect*, 26 ACC DOCKET 96 (2008); T.C., *What is the Streisand Effect*?, ECONOMIST (Apr. 16, 2013), https://www.economist.com/the-economist-explains/2013/04/15 /what-is-the-streisand-effect.

261. KRAUT & RESNICK, *supra* note 118, at 152–53.

262. IAN AYRES & JOHN BRAITHWAITE, RESPONSIVE REGULATION: TRANSCENDING THE DEREGULATION DEBATE 35–38 (Oxford Univ. Press 1992).

263. John Braithwaite, *The Essence of Responsive Regulation*, 44 U.B.C. L. REV. 475, 482 (2011). As Braithwaite cautioned, "[r]esponsive regulation asks regulators not to be dogmatic about any theory, including responsive regulation itself." *Id.* at 490.

264. GRAHAM DUKES, JOHN BRAITHWAITE & J.P. MOLONEY, PHARMACEUTICALS, CORPORATE CRIME AND PUBLIC HEALTH 289 fig.8.1 (2014).

Braithwaite explains:

> [O]ur presumption should always be to start at the base of the pyramid first. Then escalate to somewhat punitive approaches only reluctantly and only when dialogue fails. Then escalate to even more punitive approaches only when more modest sanctions fail. A regulator might escalate with a recalcitrant company from persuasion to a warning to civil penalties to criminal penalties and ultimately to corporate capital punishment-permanently revoking the company's licence to operate. . . .
>
> Strategic use of the pyramid requires the regulator to resist categorizing problems into minor matters that should be dealt with at the base of the pyramid, more serious ones that should be in the middle, and the most egregious ones for the peak of the pyramid. Even with the most serious matters—flouting legal obligations for operating a nuclear plant that risks thousands of lives, for example—we stick with the presumption that it is better to start with dialogue at the base of the pyramid.[265]

Braithwaite's pyramid of sanctions offers a potentially helpful model for remedy design: prioritize lesser sanctions initially and then progressively escalate the sanctions for recidivism. However, this model only works when the discipliners and regulated parties are in a multi-iteration game, which is not always the case for Internet services and violative users. It does not fit the situations where users seek to cause harm and never return, or where disciplined users can surreptitiously reenter the service under new identities. Both of those scenarios negate the possibility of escalated sanctions for

265. Braithwaite, *supra* note 263, at 482–83.

recidivism. Anti-recidivism techniques only work when, in fact, the punished user wants to remain in the community.

7. *Parallel Sanctions*. A service should consider how its remedies might trigger sanctions elsewhere, both judicially and extra-judicially. For example, unmasking an anonymous or pseudonymous user creates the risk of parallel consequences in other venues,[266] such as litigation, employment termination, physical violence, ostracization, reputational damage, and more. Collectively, these remedies may be disproportionate to the violation, even if the unmasking remedy itself was proportionate.

B. *Some Normative Views*

The prior subpart set out seven factors to consider as part of remedy design but did not attempt to prioritize the factors. This subpart explores some possible normative values that can help with prioritization and inform remedial design.

1. *Preserve Industry-Wide Remedial Scheme Diversity*. Due to the broad diversity of Internet services and the communities they seek to cultivate, we expect—and want—Internet services to adopt diverse content moderation remedy schemes tailored to their functions and audiences.[267] However, regulators eliminate (intentionally or not) the possibility of diverse remedial schemes when they standardize remedies across the Internet. Sometimes that makes sense, like mandatory removals for content or activity that never could be legitimate.[268] In other cases, industry-wide standardized remedies hinder the ability of Internet services to experiment with remedies or foster unique niches.

2. *Some Internet Services Have Limited Remedy Options.* Some Internet services have a limited range of technologically feasible remedy options. For example, domain name registrars cannot remove individual content items hosted by their customers; their only "removal" option is to disable the domain name,[269] which can affect legitimate content or even innocent third parties. Or, as discussed in Part II(A)(7), IAPs have limited options to control their subscribers' behavior—usually just the ability to turn Internet

266. *See* BALLON, *supra* note 46, at 37.02[2][A].

267. Schoenebeck et al., *supra* note 31, at 1295–96 ("[A] one-size-fits-all approach to online harassment may fail to support some users while privileging others. . . . [I]t is likely that a monolithic approach to governance further magnifies inequities when applied in global, cross-cultural contexts.").

268. As discussed in Part IV(A)(1), CSAM is the paradigmatic example of such content, but it is relatively unique because no additional context is required to evaluate its (il)legitimacy.

269. *E.g.*, Annemarie Bridy, *Notice and Takedown in the Domain Name System: ICANN's Ambivalent Drift into Online Content Regulation*, 74 WASH. & LEE L. REV. 1345, 1357 (2017). Because a disabled domain name functionally takes the registrant offline, even the mere threat of domain name disabling is enough to coerce most registrants to accede to any demand.

access on or off—and turning off Internet access can have disproportionate and life-changing consequences.

Services with limited remedy options are not in good positions to redress user violations.[270] These services cannot choose among highly tailored and nuanced options that may be available to other Internet services. Instead, the coarseness of the remedy options increases the odds that any remedial actions will be miscalibrated or will have adverse collateral consequences. As a result, regulators should not force these services to impose remedies for violations because the services lack appropriate tools.[271]

3. *Better Design Can Reduce Problems*. Internet services can design their services in ways that, ex ante, inhibit unwanted or violative conduct and thus reduce the need for ex post remedies.[272] This is analogous to "privacy by design" (PbD), which seeks to incorporate privacy considerations into new product and service development rather than fixing privacy violations after they've already occurred.[273] Two examples of how Internet services have experimented with ways to reduce future problems:

---

270. *See* Annemarie Bridy, *Remediating Social Media: A Layer-Conscious Approach*, 24 B.U. J. SCI. & TECH. L. 193 (2018); Joan Donovan, *Navigating the Tech Stack: When, Where and How Should We Moderate Content?*, CTR. FOR INT'L GOVERNANCE INNOVATION (Oct. 28, 2019), https://www.cigionline.org/articles/navigating-tech-stack-when-where-and-how-should-we-moderate-content.

The bluntness of account suspensions or terminations by critical vendors has occasionally generated substantial media attention, such as when:

- CloudFlare stopped providing DDOS protection to the Daily Stormer. Matthew Prince, *Why We Terminated Daily Stormer*, CLOUDFLARE BLOG (Aug. 16, 2017), https://blog.cloudflare.com/why-we-terminated-daily-stormer.
- Amazon Web Services (AWS) suspended its hosting for Parler. Parler, LLC v. Amazon Web Services, Inc., 514 F. Supp. 3d 1261 (W.D. Wash. 2021); *see also* Tony Romm & Rachel Lerman, *Amazon Suspends Parler, Taking Pro-Trump Site Offline Indefinitely*, WASH. POST (Jan. 11, 2021, 5:12 AM), https://www.washingtonpost.com/technology/2021/01/09/amazon-parler-suspension.

271. The bluntness of the proposed mandatory remedies, and the associated risks of adverse collateral consequences, was a key reason why the Stop Online Piracy Act (SOPA) posed an existential threat to the Internet. Eric Goldman, *Why I Oppose the Stop Online Piracy Act (SOPA)/E-PARASITES Act*, TECH. & MKTG. L. BLOG (Nov. 15, 2011), https://blog.ericgoldman.org/archives/2011/11/stop_online_pir.htm.

272. GILLESPIE, *supra* note 19, at 177–82 (calling it "moderation by design"); FARMER & GLASS, *supra* note 161, at 97–276; SUZOR, *supra* note 19, at 128–49; Karen Levy & Solon Barocas, *Designing Against Discrimination in Online Markets*, 32 BERKELEY TECH. L.J. 1183 (2017); Land & Hamilton, *supra* note 108, at 150 ("Platforms could be designed in ways that work to minimize the online disinhibition effect, such as through the use of cues reminding users of their shared humanity.").

273. *E.g.,* ANN CAVOUKIAN, PRIVACY BY DESIGN, THE 7 FOUNDATIONAL PRINCIPLES: IMPLEMENTATION AND MAPPING OF FAIR INFORMATION PRACTICES (2011), https://iapp.org/media/pdf/resource_center/pbd_implement_7found_principles.pdf.

- Epinions hid newly posted consumer reviews from unregistered readers until other community reviewers had rated the review as "helpful" or "very helpful."[274] This design reduced the reviews' readership, which in turn reduced the financial compensation Epinions paid for those reviews. The reduced financial incentives dissuaded many bad actors from submitting malicious reviews.
- The hyperlocal social network Nextdoor has made several design choices to discourage unwanted behavior. First, Nextdoor's "Kindness Reminder" automatically prompts users to rethink posts that looked potentially mean.[275] "In early tests in the US, 1 in 5 people who saw Kindness Reminder hit 'edit' on their comment, resulting in 20% fewer negative comments. Moreover, in areas testing Kindness Reminder, there has been a decline in how often it is prompted."[276] Second, to discourage neighbors from making crime reports based on racial profiling, Nextdoor redesigned the flow of its service so that users focused on the suspicious behavior, not a person's demographics.[277]

As the maxim goes, an ounce of prevention is worth a pound of cure.[278] Post-hoc remedies can only do so much to redress violations. Where possible, avoiding violations in the first place is preferable.

---

274. *See supra* Part III(C).

275. *About the Kindness Reminder*, NEXTDOOR HELP CTR., https://help.nextdoor.com/s/article/About-the-Kindness-Reminder (last visited Oct. 2, 2021) ("The Kindness Reminder never prevents a member from making a post, it simply aims to slow folks down in conversations that could become uncivil."). Twitter also "encourage[s] people to pause and reconsider a potentially harmful or offensive reply before they hit send." Anita Butler & Alberto Parrella, *Tweeting with Consideration*, TWITTER BLOG (May 5, 2021), https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration.html.

UGC services have been prospectively warning users about potential incivil behavior for a long time. *See* KRAUT & RESNICK, *supra* note 118, at 150–51. For example, in 2000, the email software program Eudora incorporated a feature called "MoodWatch" that alerted users when it detected they were writing a "flame" email. *Qualcomm's Eudora 5.0 Spices up the Email Experience with Hot, New Time-Saving Tools to Keep People Connected*, QUALCOMM (Sept. 11, 2000), https://www.qualcomm.com/news/releases/2000/09/11/qualcomms-eudora-50-spices-email-experience-hot-new-time-saving-tools-keep; *see also* Jim W. Ko, *The Fourth Amendment and the Wiretap Act Fail to Protect Against Random ISP Monitoring of E-mails for the Purpose of Assisting Law Enforcement*, 22 J. MARSHALL J. COMPUT. & INFO. L. 493, 496 (2004).

276. Tatyana Mamut, *Announcing Our New Feature to Promote Kindness in Neighborhoods*, NEXTDOOR BLOG (Sept. 18, 2019), https://blog.nextdoor.com/2019/09/18/announcing-our-new-feature-to-promote-kindness-in-neighborhoods; *see also* Land & Hamilton, *supra* note 108, at 150 (noting the benefits of adding "friction" into user interfaces).

277. *Nextdoor's Approach*, NEXTDOOR, https://go.us.nextdoor.com/safety/preventing-profiling-approach (last visited Oct. 2, 2021).

278. A phrase attributed to Benjamin Franklin.

User education and socialization can also discourage violations from occurring.[279] New user onboarding is an optimal time to socialize them about community norms,[280] though retaining shared community norms becomes harder as the community size grows.

Wikipedia provides a useful case study. By design, Wikipedia makes it trivially easy for anyone to edit articles,[281] but it also has an elaborate and baroque socialization and acculturation process for converting casual readers into highly engaged "Wikipedians."[282] The process screens out many qualified contributors, but the editors who remain engaged become well-socialized in Wikipedia's norms and expectations.

4. *Private Remedies Are (Usually) Preferable to Judicial Remedies.* Courts are typically the gold standard for adjudicating the legitimacy of content or conduct. Courts have a high degree of expertise and accuracy in admitting evidence and applying the applicable law to that evidence, and they follow procedures that inspire confidence and trust in their outcomes.

Nevertheless, courts may not be well-positioned to determine remedies for online violations because their adjudications: (1) take a long time, with harm possibly accruing during the pendency; (2) often cost more than the social value of the dispute; and (3) may be limited by jurisdictional problems reaching the disputants. Furthermore, court proceedings increase the risk of the Streisand Effect,[283] which can conflict with other remedies.[284]

---

279. *See* Grimmelmann, *supra* note 12, at 61–63; WAGNER ET AL., *supra* note 105, at 24–25; Jhaver et al., *supra* note 12.

280. KRAUT & RESNICK, *supra* note 118, ch. 4 (discussing approaches to regulating user behavior).

281. *See Help:Editing*, WIKIPEDIA, https://en.wikipedia.org/wiki/Help:Editing (last visited Oct. 2, 2021).

282. *See generally* Goldman, *supra* note 178, at 167–69.

283. *See* O'Kroley v. Fastcase, Inc., 831 F.3d 352, 356 (6th Cir. 2016):

> In most respects, O'Kroley didn't accomplish much in suing Google and the other defendants. He didn't win. He didn't collect a dime. And the search result about "indecency with a child" remains publicly available. All is not lost, however. Since filing the case, Google users searching for "Colin O'Kroley" no longer see the objectionable search result at the top of the list. Now the top hits all involve this case (there is even a Wikipedia entry on it). So: Even assuming two premises of this lawsuit are true—that there are Internet users other than Colin O'Kroley searching "Colin O'Kroley" and that they look only at the Google previews rather than clicking on and exploring the links—it's not likely that anyone will ever see the offending listing at the root of this lawsuit. Each age has its own form of self-help.

284. For example, to avoid giving counterproductive publicity to information that must be made more obscure, Google cannot notify sites that it is delisting their pages in response to "right to be forgotten" (RTBF) demands. *See The Swedish Data Protection Authority Imposes Administrative Fine on Google,* EUR. DATA PROT. BD. (Mar. 11, 2020), https://edpb.europa.eu/news/national-news/2020/swedish-data-protection-authority-imposes-administrative-fine-google_en.

When Internet services decide for themselves to impose remedies for user violations, they avoid most of the courts' limitations but create different challenges. Internet services may not invest adequately in determining an appropriate remedy because customized/"artisanal" remedies may be cost- and time-prohibitive. Furthermore, many Internet services have low incentives to hear both sides or follow due process. Finally, Internet services choose remedies that maximize their interests, not the best interests of the affected user, the community, or society generally.[285]

Despite those limitations, Internet services are best positioned to understand their communities and the tricky value tradeoffs unique to their communities.[286] They also face marketplace consequences for miscalibrating their remedies.[287] For these reasons, in most situations, we should prefer that Internet services, not courts, decide the appropriate remedies.

5. *Remedies Should Be Necessary and Proportionate*. Human rights law has long dictated that remedies should be imposed only as necessary to achieve a legitimate aim and proportionate to the aim.[288] This principle can extend to content moderation remedies. In general, Internet services should impose remedies only as necessary to achieve a legitimate remedial outcome and as proportionate to the violation's nature and severity.[289] This also embraces the spirit of the First Amendment jurisprudential concept that speech restrictions should be the "least restrictive means" available to redress the government's objectives.[290]

Two specific ways to operationalize this principle:

(1) Where possible, impose remedies against individual content items rather than accounts.[291] This reduces the risk of unexpected collateral damage caused by account restrictions.

---

285. *See supra* Part IV(A)(5).

286. *Cf.* Golden, *supra* note 225, at 512 (discussing the "devolution" principle, which is to place "considerable discretion in the hands of private parties and government actors nearest to the facts of individual cases").

287. Eric Goldman, *Regulating Reputation*, *in* THE REPUTATION SOCIETY: HOW ONLINE OPINIONS ARE RESHAPING THE OFFLINE WORLD 51 (Hassan Masum & Mark Tovey eds., 2012).

288. *See*, *e.g.*, NECESSARY & PROPORTIONATE, INTERNATIONAL PRINCIPLES ON THE APPLICATION OF HUMAN RIGHTS TO COMMUNICATIONS SURVEILLANCE (2014), https://necessaryandproportionate.org/files/en_principles_2014.pdf.

289. *Oversight Bd. Decision*, *supra* note 144 (stating that content restrictions "must be necessary and proportionate to the risk of harm"); Douek, *supra* note 33, at 785–89.

290. 1 SMOLLA & NIMMER ON FREEDOM OF SPEECH §§ 4:21, 4:25 (2020), Westlaw FREESPEECH; *see also Oversight Bd. Decision*, *supra* note 144, at ("Facebook should use less restrictive measures to address potentially harmful speech and protect the rights of others before resorting to content removal and account restriction."); Douek, *supra* note 33, at 825 ("[P]roportionality requires that enforcement be the *least restrictive means*.").

291. *See*, *e.g.*, YouTube Report, *supra* note 22, at 15 ("[I]f an individual app infringes on our policies, we typically take action on that specific app rather than sanctioning the account of the developer.").

(2) Where possible, impose "local" rather than "global" remedies. For example, with respect to content published globally, the service should implement remedies only in the country(ies) where the content actually violates the local law, not globally.[292] Similarly, services might impose remedies that affect the experience of only a segment of their communities, such as age-gating content to reduce its exposure to children while preserving it for adults,[293] or Epinions' approach of displaying lowly rated reviews only to registered users instead of its entire audience.

6. *Prefer Remedies That Empower Readers*. Where possible, it is preferable to empower users to decide what they want to see, rather than imposing remedies that universally affect all users.[294]

User-controlled filters have a venerable tradition in online spaces. From the earliest days, Internet services have provided "mute" functionality that allowed one user to avoid the content of another user. The 1990s virtual world LambdaMOO called its mute feature "@gag,"[295] the pioneering online service The WELL deployed "bozo filters,"[296] and USENET users could use "kill files" to block incoming messages from specified individuals.[297] Modern examples include YouTube's Restricted Mode[298] and Block Party's anti-harassment filters.[299]

Ideally, services will compete with each other to provide the most user-beneficial filtering option, thus expanding user choice among filters and spurring new innovation in filtering approaches.[300] This has been the basic

---

This is consistent with the "ABC framework" of distinguishing among "actors, behavior, and content." *See* TRANSATLANTIC WORKING GRP., *supra* note 24, at 18–21 (2020).

292. MANILA PRINCIPLES, *supra* note 58, at 4; *see also Oversight Bd. Decision*, *supra* note 144, at 29 (noting that least restrictive measures include "developing effective mechanisms to avoid amplifying speech that poses risks of imminent violence, discrimination, or other lawless action, where possible and proportionate, rather than banning the speech outright").

293. *See Your Content & Restricted Mode*, YOUTUBE HELP, https://support.google.com/youtube/answer/7354993?hl=en (last visited Oct. 2, 2021).

294. *See* Ashcroft v. ACLU, 542 U.S. 656 (2004).

295. Mnookin, *supra* note 35 ("In LambdaMOO, any player can 'gag' any other player (or object); issuing the '@gag' command prevents the gagged player's words from appearing on the issuer's screen."). LambdaMOO also provided a "refuse" feature that allowed a player to block all incoming messages from another player. *Id.*

296. Howard Rheingold, *Bozo Filters*, WIRED (Jan. 1, 1993, 12:00 PM), https://www.wired.com/1993/01/bozo-filters.

297. ADAM GAFFIN & JORG HEITKOTTER, EFF'S (EXTENDED) GUIDE TO THE INTERNET 77 (1994), https://archive.org/stream/B-001-004-387/eegtti-2.3-a5_djvu.txt. Adding someone to a kill file was called "plonking." *Definitions for Plonk*, DEFINITIONS.NET, https://www.definitions.net/definition/Plonk (last visited Nov. 18, 2021).

298. *Your Content & Restricted Mode*, supra note 293.

299. BLOCK PARTY, https://www.blockpartyapp.com (last visited Oct. 2, 2021).

300. *Cf.* TIMOTHY GARTON ASH ET AL., REUTERS INST. & UNIV. OF OXFORD, GLASNOST! NINE WAYS FACEBOOK CAN MAKE ITSELF A BETTER FORUM FOR FREE SPEECH AND DEMOCRACY 16–17 (2019) (discussing various alternative filters that Facebook could offer to its users).

premise of email anti-spam filters; each email service adopts its own, and email customers can choose between services in part based on the efficacy of the services' anti-spam filters.

It is possible to implement competition among filters on an even grander scale. Mike Masnick has proposed that social media services reconfigure themselves into "protocols, not platforms."[301] The idea is to decouple content collection from its publication. Social media services would still collect and publish user content, as they have always done, but the services would also make the corpus of collected content available to other services, who could then republish it themselves. Each service could determine and apply their own editorial standards for the corpus, including content ranking/ordering and content moderation. This transition would increase competition among the rival services to provide the best experience for users.[302] In particular, services could adopt heterogeneous approaches to remedies for violations and use their remedial schemes as points of competitive differentiation. Though this approach remains theoretical, Twitter is actively exploring a protocols-not-platforms implementation via its "Blue Sky Project."[303]

All technological filters inevitably create the risk of "filter bubbles,"[304] where users choose to see only what reinforces their preexisting knowledge and biases. Filter bubbles are a real concern, as such closed-loop information systems thwart users' exposure to the realities faced by others. On the other hand, the tradeoff will often devolve into a choice between content being categorically suppressed and content being consumed only among those in a filter bubble. Neither is ideal, but it is not clear that we should prefer categorical suppression.

Another user-empowerment approach is to provide users with more disclosures about possibly violative material, such as legends, labels, or warnings. This is the age-old approach of counterspeech and contextualization, and it is the preferred approach for advocates of the "marketplace of ideas." In theory, these disclosures improve readers' choices about what content to consume and how credible it is.

---

301. Mike Masnick, *Protocols, Not Platforms: A Technological Approach to Free Speech*, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech.

302. *Id.* ("[W]e can let a million content moderation systems approach the same general corpus of content—each taking an entirely different approach—and see which ones work best.") Note this assumes regulators do not force the collecting service to over-remove content pre-dissemination. If the collecting service is legally obligated to disseminate only non-violative content, it limits the capacity for remedial competition among services.

303. Jack Dorsey (@Jack), TWITTER (Jan. 13, 2021, 7:16 PM), https://twitter.com/jack/status/1349510769268850690. Services deploying related concepts include diaspora* and Mastadon. WAGNER ET AL., *supra* note 105, at 16, 18.

304. ELI PARISER, THE FILTER BUBBLE: HOW THE NEW PERSONALIZED WEB IS CHANGING WHAT WE READ AND HOW WE THINK (2011).

Unfortunately, in practice, disclosures have well-known efficacy limits.[305] It is hard to educate consumers enough to make well-informed decisions about anything. Disclosures also can actively mislead users or counterproductively induce reader reactance.[306] Despite these significant problems, disclosures retain an important place in the remedial toolkit because of their user empowerment.

C. *Implications for "Platform" Transparency*

Regulators, civil society, and individual consumers are demanding greater transparency from "platforms."[307] Transparency plays a major role in the "platform governance" academic discussion, and there is an active academic discourse about "transparency." This Article presents some implications for those transparency discussions.

Historically, many Internet services' transparency reports have disclosed removals (or similar actions, like suspension). For example, the Santa Clara Principles says signatories should "publish the numbers of posts removed and accounts permanently or temporarily suspended due to violations of their content guidelines."[308] This essentially codified the binary dichotomy into the transparency reports.

An expanded toolkit of content moderation remedies creates new challenges for transparency disclosures. What level of granularity about remedies should be disclosed, and at what cost?[309] If an Internet service deploys a dozen different remedial options, providing granular disclosures about each remedy would increase the report's complexity, and the associated data gathering, by twelve-fold. If an Internet service provided even more granular disclosures about the remedies, such as remedy imposition by type of rule violation, the complexity of the disclosures—and the Internet service's backend systems needed to produce it—grows exponentially.

---

305. *E.g.*, OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE (2014).

306. Fact-checking "may cause people to 'double-down' on their incorrect beliefs, producing a backlash effect." Alice E. Marwick, *Why Do People Share Fake News? A Sociotechnical Model of Media Effects*, 2 GEO. L. TECH. REV. 474, 475 (2018); *see id.* at 508; *see also* Land & Hamilton, *supra* note 108, at 152 (explaining some limits of counter-speech).

307. *E.g.*, *Removals Under the Network Enforcement Law*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/netzdg/overview?hl=en (last visited Oct. 26, 2021); *The Digital Services Act: Ensuring a Safe and Accountable Online Environment*, EUR. COMM'N, https://ec.europa.eu/info/strategy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en (last visited Oct. 26, 2021).

308. *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 67.

309. *See generally Oversight Bd. Decision*, *supra* note 144 (discussing its expectations for Facebook's disclosures about remedies).

Alternatively, Internet services could simply disclose the number of times they took any remedial action, without detailing the specific remedy.[310] A more generic disclosure like this would be less costly but also less insightful; removal has different consequences than non-removal remedies (such as reduced virality or visibility), and those differences may be significant enough to matter to the report's implications.[311]

Thus, regulators should consider how increased remedy options may affect any mandated transparency obligations, including the understandability of the reports and the production costs.

With respect to production costs, Internet services may possess the relevant data about usage of different remedies and the associated consequences. However, many services have not built the tools needed to report or analyze the data.[312] Also, building those tools to regulators' specifications may be cost-prohibitive. Furthermore, sharing the reports (to regulators or the public) could raise privacy concerns.

These costs may be outweighed by the benefits of greater transparency obligations for Internet services' remedies. However, regulators should clearly identify: who is the audience for the produced data; what decision(s) that audience will be making based on the data; and how the produced data will improve their decisions.[313] Without such clarity, it is not clear the data will produce any benefits at all.

Finally, the decision of which metrics to track and report subsumes some important considerations about what values we should prioritize. Remedial schemes can encode multiple values that point in different directions. To the extent that remedial design unavoidably involves tradeoffs between competing values, picking single metrics to optimize can

---

310. The 2020 version of the PACT Act would have required Internet services to disclose "the number of instances in which the interactive computer service provider took action with respect to illegal content, illegal activity, or known potentially policy-violating content . . . including content removal, content demonetization, content deprioritization, appending content with an assessment, account suspension, account removal, or any other action taken in accordance with the acceptable use policy of the provider . . . ." Platform Accountability and Consumer Transparency Act, S. 4066, 116th Cong. § 5(d)(2)(B) (2020). Though not part of the proposed language, requiring Internet services to disclose which action they took per item/account would dramatically increase the costs of the transparency requirement and its potential for errors.

311. This appears to be another example of the accuracy-simplicity tradeoff. *E.g.*, Enriqueta Aragones et al., *Accuracy vs. Simplicity: A Complex Trade-Off* (2003), https://www.sas.upenn.edu/~apostlew/paper/pdf/AGPS.pdf.

312. External constituencies might be able to generate useful insights through application programming interfaces (APIs) (if the services make them available) or by data scraping (though scraping may be legally dubious). *See* Eric Goldman, *The Constitutionality of Mandated Editorial Transparency*, 73 HASTINGS L.J. (forthcoming 2022).

313. *Cf.* ARCHON FUNG, MARY GRAHAM, & DAVID WEIL, FULL DISCLOSURE: THE PERILS AND PROMISE OF TRANSPARENCY (2009) (discussing the history of transparency policies in the US and relating them to market forces and information regulation).

be misleading or even harmful. At minimum, any tracked metrics should not obscure the unavoidable tradeoffs from any remedial scheme.

CONCLUSION

After two decades of "techno-optimism"[314] about the Internet and its potential, the pendulum has swung sharply in the opposite direction. There is widespread pessimism about the Internet and its effects on society.[315] This has dramatically ramped up regulator—and popular—support for "crackdowns" on bad Internet content and actors, even if those crackdowns constitute censorship[316] or will cause massive collateral damage.[317] These dynamics have created a seemingly unstoppable push for structural reforms to the Internet, regardless of policy merit.

Similarly, content moderation remedies have become partisan.[318] As an oversimplification, liberals want more user content permanently removed, and conservatives want more user content left completely untouched. In theory, this Article offers an alternative way of thinking about content moderation that might defuse the partisan tension. More likely, this Article is so far outside the current Overton Window[319] that it will not satisfy partisans on either side.

Content moderation is hard. It is not possible to moderate content in a way that pleases everyone.[320] That makes regulatory interventions into the content moderation process particularly dangerous. The interventions have high risks of increasing Internet services' costs while still leaving everyone dissatisfied.

This Article shows how a diversity of content moderation remedies offer interesting and underappreciated options to help Internet services better serve their communities and balance many competing interests. These

---

314. *E.g.*, Margaret O'Mara, *The Church of Techno-Optimism*, N.Y. TIMES (Sept. 28, 2019), https://www.nytimes.com/2019/09/28/opinion/sunday/silicon-valley-techno-optimism.html.

315. *See*, *e.g.*, Aaron Smith, *Declining Majority of Online Adults Say the Internet Has Been Good for Society*, PEW RESEARCH CTR. (Apr. 30, 2018), https://www.pewresearch.org/internet/2018/04/30/declining-majority-of-online-adults-say-the-internet-has-been-good-for-society; *see generally* THE SOCIAL DILEMMA (Netflix 2020).

316. *E.g.*, *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) (Austl.); Stop the Censorship Act, H.R. 7808, 116th Cong. (2020); Ending Support for Internet Censorship Act, S. 1914, 116th Cong. (2019).

317. Directive 2019/790, of the European Parliament and the Council of 17 April 2019, on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9 /EC and 2001/29/EC, art. 17, 2019 O.J. (L 130) 92, 119-121.

318. *E.g.*, Alayna Treene, Margaret Harding McGill & Ashley Hold, *GOP Plots Payback for Deplatforming Trump*, AXIOS (Jan. 17, 2021), https://www.axios.com /republicans-tech-trump-5a85a2dc-8360-4d29-87b1-1da7cc9abfed.html.

319. *E.g.*, Maggie Astor, *How the Politically Unthinkable Can Become Mainstream*, N.Y. TIMES (Feb. 26, 2019), https://www.nytimes.com/2019/02/26/us/politics/overton-window-democrats.html (discussing the origins and implications of the Overton Window).

320. Goldman, *supra* note 255.

options become politically relevant only if regulators exercise self-restraint and do not hard-code remedies that strip Internet services of remedial discretion. The current regulatory maelstrom, with the seemingly unshakable and singular focus of permanently ending the era of user-generated content,[321] reduces the odds that we will realize the benefits of this underexplored toolkit.

321. Goldman, *UK Online Harms*, *supra* note 21, at 360–62.