COLUMBIA LAW REVIEW

VOL. 121 APRIL 2021 NO. 3

ARTICLES

GOVERNING ONLINE SPEECH: FROM "POSTS-AS-TRUMPS" TO PROPORTIONALITY AND PROBABILITY

*Evelyn Douek**

Online speech governance stands at an inflection point. The state of emergency that platforms invoked during the COVID-19 pandemic is subsiding, and lawmakers are poised to transform the regulatory landscape. What emerges from this moment will shape the most important channels for communication in the modern era and have profound consequences for individuals, societies, and democratic governance. Tracing the path to this point illuminates the tasks that the institutions created during this transformation must be designed to do. This history shows that where online speech governance was once dominated by the First Amendment tradition's categorical and individualistic approach to adjudicating speech conflicts, that approach became strained, and online speech governance now revolves around two other principles: proportionality and probability. Proportionality requires no longer focusing on the speech interest in an individual post alone, but also taking account of other societal interests that can justify proportionate limitations on content. But the unfathomable scale of online speech makes enforcement of rules only ever a matter of probability: Content moderation will always involve error, and so the pertinent questions are what error rates are reasonable and which kinds of errors should be preferred. Platforms' actions during the pandemic have thrown into stark relief the centrality of these principles to online speech governance and also how undertheorized they remain. This Article reviews the causes of this shift from a "posts-as-trumps" approach to online speech governance to one of systemic balancing and what this new era of content moderation entails for platforms and their regulators.

* Lecturer on Law & S.J.D. Candidate, Harvard Law School; Affiliate, Berkman Klein Center for Internet & Society. Many thanks to Jack Balkin, Ros Dixon, Jack Goldsmith, Vicki Jackson, Martha Minow, Rory Van Loo, Jonathan Zittrain, and discussants at Harvard Law School's S.J.D. Colloquium, Yale I.S.P.'s Ideas Lunch, and Harvard's Berkman Klein Center for Internet & Society for extremely helpful comments and conversations. I am also very grateful to Tarek Roshdy and the editors of the *Columbia Law Review* for their excellent work and general good humor. Errors are categorically (not merely proportionately or probably) my own.

INTRODUCTION 761
I. THE OLD AND NEW PRECEPTS OF ONLINE SPEECH GOVERNANCE 769
A. Posts-As-Trumps 770
1. The Role of Formal Law 770
2. Cultural and Legal Familiarity 771
3. Methodological Exceptionalism and Professed Humility 772
B. The Rise of Proportionality 776
1. Reasons Behind the Rise 776
2. The Pros of Proportionality 785
C. The Unavoidability of Probability 789
1. Content Moderation Is Impossible 791
2. Tools and Systems 792
3. Probability as Pragmatism 798
D. The Pandemic 800
1. The Pandemic and Proportionality 800
2. The Pandemic and Probabilities 802
II. ASKING THE RIGHTS QUESTIONS 804
A. Incommensurability 804
B. Getting Rights Wrong 808
III. RECALIBRATING 813
A. Against Speech Rights Deflation 814
B. Addressing Systemic Balancing's Current Deficits 819
1. Systemic Balancing for Platforms 821
2. Systemic Balancing for Regulators 826
C. Returning from the Not-So-Exceptional State of Exception 830
CONCLUSION 833

Twitter being abused to instill fear, to silence your voice*, or to undermine individual safety, is unacceptable.*

— @TwitterSafety, October 3, 2020[1]

A commitment to expression is paramount, but we recognize the internet creates new and increased opportunities for abuse. For these reasons, when we limit expression we do it in service of one or more of the following values: Authenticity . . . Safety . . . Privacy . . . Dignity.

— Monika Bickert, Facebook Vice President, Global Policy Management, September 12, 2019[2]

I have to admit that I've struggled with balancing my values as an American, and around free speech and free expression, with my values and the company's values around common human decency.

— Steven Huffman, Reddit CEO, June 29, 2020[3]

INTRODUCTION

On March 6, 2020, Facebook announced it was banning ads for medical face masks across its platforms to prevent people from trying to exploit the COVID-19 public health emergency for commercial gain.[4] A month later, the *New York Times* reported that Facebook's ban was hampering volunteer efforts to create handsewn masks for medical professionals, as Facebook's automated content moderation systems over-enforced the mask-ad ban.[5] At the same time, *BuzzFeed* reported, Facebook was still profiting off scammers running mask ads not caught by those same systems.[6] On June 10, 2020, Facebook noted that authorities' guidance on

1. Twitter Safety (@TwitterSafety), Twitter (Oct. 3, 2020), https://twitter.com/TwitterSafety/status/1312498519094091779 (on file with the *Columbia Law Review*) (emphasis added).

2. Monika Bickert, Updating the Values that Inform Our Community Standards, Facebook: Newsroom (Sept. 12, 2019), https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards [https://perma.cc/8MF6-N8WV] [hereinafter Bickert, Updating the Values].

3. Casey Newton, Reddit Bans r/The_Donald and r/ChapoTrapHouse as Part of a Major Expansion of Its Rules, Verge (June 29, 2020), https://www.theverge.com/2020/6/29/21304947/reddit-ban-subreddits-the-donald-chapo-trap-house-new-content-policy-rules (on file with the *Columbia Law Review*) [hereinafter Newton, Reddit Bans].

4. Guy Rosen, An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19, Facebook: Newsroom (Apr. 16, 2020), https://about.fb.com/news/2020/04/covid-19-misinfo-update [https://perma.cc/6JMQ-GAV9] [hereinafter Rosen, Facebook COVID-19 Update] (last updated May 12, 2020).

5. Mike Isaac, Facebook Hampers Do-It-Yourself Mask Efforts, N.Y. Times (Apr. 5, 2020), https://www.nytimes.com/2020/04/05/technology/coronavirus-facebook-masks.html (on file with the *Columbia Law Review*).

6. Craig Silverman, Facebook Banned Mask Ads. They're Still Running., BuzzFeed News (May 13, 2020), https://www.buzzfeednews.com/article/craigsilverman/facebook-

wearing masks had "evolved" since the start of the pandemic, and the ban would be scaled back to permit promotion of nonmedical masks.[7] This was well after many experts had begun recommending masks,[8] but only shortly after the WHO changed its guidance.[9]

This mask-ad ban example is a microcosm of the key challenges of content moderation on the largest social media platforms. The scale at which these platforms operate means mistakes in enforcing any rule are inevitable: It will always be possible to find examples of both false positives (taking down volunteer mask makers) and false negatives (mask ads being approved to run on the site). In writing and enforcing a mask-ad ban, then, the issue is not simply whether such a ban is good in principle but also how to make trade-offs between speed, nuance, accuracy, and over- or under-enforcement. Whether to enact a ban in the first place is fraught too. Platforms justified their unusually interventionist approach to false information in the context of the COVID-19 pandemic in part because there were more clear-cut "authoritative" sources of information, such as the WHO, to which they could defer.[10] So what should platforms do when, as in the case of masks, those authorities increasingly contradict scientific consensus, or in other contexts where such clearly identifiable authorities do not exist?

There are no easy answers, but moving the project of online speech governance forward requires asking the right questions. Instead of thinking about content moderation through an individualistic lens typical of constitutional jurisprudence, platforms, regulators, and the public at large need to recognize that the First Amendment–inflected approach to online speech governance that dominated the early internet no longer holds. Instead, platforms are now firmly in the business of balancing societal interests and choosing between error costs on a systemic basis. This Article shows that these choices are endemic to every aspect of modern online speech governance and suggests that this requires a recalibration of our

---

mask-ads-ban-zestads-coronavirus [https://perma.cc/8Q8E-H7WW] [hereinafter Silverman, Facebook Mask-Ad Ban].

7. Rob Leathern, Allowing the Promotion of Non-Medical Masks on Facebook, Facebook for Business (June 10, 2020), https://www.facebook.com/business/news/allowing-the-promotion-of-non-medical-masks-on-facebook [https://perma.cc/R788-X6R7] (last updated Aug. 19, 2020).

8. See Zeynep Tufekci, Jeremy Howard & Trisha Greenhalgh, The Real Reason to Wear a Mask, Atlantic (Apr. 22, 2020), https://www.theatlantic.com/health/archive/2020/04/dont-wear-mask-yourself/610336 (on file with the *Columbia Law Review*) (discussing advice from experts to wear masks).

9. Sarah Boseley, WHO Advises Public to Wear Face Masks When Unable to Distance, Guardian (June 5, 2020), https://www.theguardian.com/world/2020/jun/05/who-changes-advice-medical-grade-masks-over-60s [https://perma.cc/JGV5-7JPL].

10. See, e.g., Press Release, Facebook, Facebook Press Call 17 (Mar. 18. 2020), https://about.fb.com/wp-content/uploads/2020/03/March-18-2020-Press-Call-Transcript.pdf [https://perma.cc/3XCR-PD8M] [hereinafter Facebook Press Call] ("[T]he WHO for example . . . have broad trust and a government mandate on [COVID-19] in a way that in other domains there just (isn't) something like that.").

understanding of content moderation—the systems for writing and enforcing the rules for what social media platforms allow on their services.

This project of recalibration is urgent: Online speech governance stands at an inflection point. Lawmakers in the United States and abroad are poised to radically transform the existing legal landscape (and in some cases have already started doing so);[11] platforms are both trying to get ahead of these developments and playing catch-up to societal demands for more responsible content moderation through self-regulatory innovations and reforms.[12] Content moderation entered a "state of emergency" during the COVID-19 pandemic,[13] but the emergency is starting to subside. The governance institutions that emerge from this upheaval will define the future of online speech and, with it, modern public discourse.

Designing these institutions requires understanding the evolution of platform governance so far and what this reveals about the underlying dynamics of content moderation. That story shows that content moderation on major platforms, once dominated by a categorical and individualistic conception of online speech rights, is now crafted around two different precepts: proportionality and probability. That is, content moderation is a question of *systemic balancing*: Rules are written to encompass multiple interests, not just individual speech rights, and with awareness of the error rates inherent in enforcing any rule at the truly staggering scale of major platforms.

Recognizing this shift illuminates the nature of adjudication required.[14] Decisions centered around proportionality and probability are different in kind. Proportionality necessitates intrusions on rights being justified, and greater intrusions having stronger justifications.[15] In constitutional systems, proportionality takes various doctrinal forms but always involves a balancing test that requires the decisionmaker to balance

---

11. See Spandana Singh, Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content 9–11 (2019), https://d1y8sb8igg2f8e.cloudfront.net/documents/Everything_in_Moderation_2019-07-15_142127_tq36vr4.pdf [https://perma.cc/96D9-CUW9].

12. See, e.g., Evelyn Douek, "What Kind of Oversight Board Have You Given Us?", U. Chi. L. Rev. Online (May 11, 2020), https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek [https://perma.cc/V329-H8Y8] (exploring the design and potential of the Facebook Oversight Board, an "unprecedented experiment in content moderation governance").

13. Evelyn Douek, The Internet's Titans Make a Power Grab, Atlantic (Apr. 18, 2020), https://www.theatlantic.com/ideas/archive/2020/04/pandemic-facebook-and-twitter-grab-more-power/610213 (on file with the *Columbia Law Review*) [hereinafter Douek, The Internet's Titans].

14. See Alec Stone Sweet & Jud Mathews, Proportionality Balancing and Constitutional Governance: A Comparative and Global Approach 13–14 (2019) [hereinafter Stone Sweet & Mathews, Proportionality Balancing and Constitutional Governance] (noting that different conceptions of rights "produce different approaches to rights adjudication").

15. Vicki C. Jackson, Constitutional Law in an Age of Proportionality, 124 Yale L.J. 3094, 3117–18 (2015).

societal interests against individual rights.[16] This emphasis on justification and balancing therefore takes the decisionmaker from being a mere "taxonomist[]"[17] (categorizing types of content) to grocer (placing competing rights and interests on a scale and weighing them against each other)[18] or epidemiologist (assessing risks to public health).[19] This task requires much greater transparency of reasoning.

Meanwhile, a probabilistic conception of online speech acknowledges that enforcement of the rules made as a result of this balancing will never be perfect, and so governance systems should take into account the inevitability of error and choose what kinds of errors to prefer.[20] The conscious acceptance of the fact that getting speech determinations wrong in some percentage of cases is inherent in online speech governance requires being much more candid about error rates, which can allow for the calibration of rulemaking to the practical realities of enforcement.

The arrival of this new era in online speech governance is increasingly apparent, even if usually only implicitly acknowledged. Professor Jonathan Zittrain has observed a move from a "rights" era of online governance to a "public health" one that requires weighing risks and benefits of speech.[21] Professor Tim Wu describes the "open and free" speech ideal of the first twenty years of the internet changing "decisively" to a "widespread if not universal emphasis among the major platforms . . . on creating 'healthy' and 'safe' speech environments online."[22] Contract for the Web, founded by Tim Berners-Lee, inventor of the World Wide Web, has called for companies to address the "risks created by their technologies," including their online content, alongside their benefits.[23] It is now fairly common to hear calls that "[c]ontent moderation on social platforms needs to *balance* the impact on society with the individual rights of speakers and the right

16. Richard H. Fallon, Jr., Strict Judicial Scrutiny, 54 UCLA L. Rev. 1267, 1296 (2007).

17. Kathleen M. Sullivan, Post-Liberal Judging: The Roles of Categorization and Balancing, 63 U. Colo. L. Rev. 293, 293 (1992).

18. Id. at 293–94.

19. John Bowers & Jonathan Zittrain, Answering Impossible Questions: Content Governance in an Age of Disinformation, Harv. Kennedy Sch. Misinfo. Rev., Jan. 2020, at 1, 4–5.

20. See infra section II.B.

21. See Jonathan Zittrain, Three Eras of Digital Governance 1 (Sept. 15, 2019) (unpublished manuscript), https://www.ssrn.com/abstract=3458435 (on file with the *Columbia Law Review*) [hereinafter Zittrain, Three Eras].

22. Tim Wu, Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems, 119 Colum. L. Rev. 2001, 2009 (2019) [hereinafter Wu, Will Artificial Intelligence Eat the Law].

23. Principle 6: Develop Technologies that Support the Best in Humanity and Challenge the Worst, Cont. for the Web, https://contractfortheweb.org/principles/principle-6-develop-technologies-that-support-the-best-in-humanity-and-challenge-the-worst [https://perma.cc/L4F2-7DNJ] (last visited Oct. 23, 2020).

for people to consume the content of their choice."[24] A civil rights audit of Facebook admonished the company for still taking an unduly "selective view of free expression as Facebook's most cherished value" without accounting for impacts on other rights.[25]

Facebook's update to the "values" that inform its Community Standards is perhaps the starkest example of the dominance of this new paradigm.[26] Where once Facebook emphasized connecting people,[27] it now acknowledges that voice should be limited for reasons of authenticity, safety, privacy, and dignity.[28] As a result, "Although the Community Standards do not explicitly reference proportionality, the method described . . . invokes some elements of a traditional proportionality test."[29] Similarly, Twitter CEO Jack Dorsey has acknowledged that Twitter's early rules "likely over-rotated on one value" and the platform would now root its rules in "human rights law,"[30] which includes a proportionality test.[31]

Similarly, there has been increasing acknowledgment that enforcement of rules will never be perfect.[32] That is, content moderation will always be a matter of probability. Tech companies and commentators accept that the volume of speech made tractable and, therefore, in some sense governable as a result of that speech migrating online makes it

---

24. Mathew Ingram, Former Facebook Security Chief Alex Stamos Talks About Political Advertising, Galley by CJR, https://galley.cjr.org/public/conversations/-LyjQOoPX4-yK-H78Mw6 (on file with the *Columbia Law Review*) (last visited Oct. 23, 2020) (emphasis added).

25. Facebook's Civil Rights Audit—Final Report 9 (2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/E5SX-CPWK].

26. See Bickert, Updating the Values, supra note 2.

27. See, e.g., Note from Mark Zuckerberg, Facebook: Newsroom (Apr. 27, 2016), https://newsroom.fb.com/news/2016/04/marknote [https://perma.cc/9QAG-ESGL] (stating Facebook's mission as "mak[ing] the world more open and connected").

28. See Evelyn Douek, Why Facebook's "Values" Update Matters, Lawfare (Sept. 16, 2019), https://www.lawfareblog.com/why-facebooks-values-update-matters [https://perma.cc/3ZDK-VXPK] [hereinafter Douek, Why Facebook's Update Matters].

29. Matthias C. Kettemann & Wolfgang Schulz, Setting Rules for 2.7 Billion: A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study 20 (Hans-Bredow-Institut Working Paper No. 1, 2020), https://www.hans-bredow-institut.de/uploads/media/default/cms/media/k0gjxdi_AP_WiP001InsideFacebook.pdf [https://perma.cc/8EB5-XSVS].

30. Jack Dorsey (@jack), Twitter (Aug. 10, 2018), https://twitter.com/jack/status/1027962500438843397 (on file with the *Columbia Law Review*).

31. See, e.g., David Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, at 4, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) (identifying "proportionality" as one of the requirements for "State limitations on freedom of expression").

32. See, e.g., Monika Bickert, Facebook, Charting a Way Forward: Online Content Regulation 7 (2020), https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf [https://perma.cc/2E37-RRRH] [hereinafter Bickert, Charting a Way Forward] ("[I]nternet companies' enforcement of content standards will always be imperfect.").

unrealistic to expect rules to be applied correctly in every case.[33] The discourse is (slowly) shifting from simple exhortations to "do better" and "nerd harder,"[34] to more nuanced conversations about how to align incentives so that all relevant interests are balanced and unavoidable error costs are not disproportionately assigned in any direction.[35]

Content moderation practices during the COVID-19 pandemic have epitomized this new paradigm,[36] throwing into sharp relief the interest balancing and error choices that platforms make. Platforms cracked down on misinformation in an unprecedented fashion because the harms were judged to be especially great.[37] They did this despite acknowledging that circumstances meant there would be higher error rates than normal because the costs of moderating inadequately were less than the costs of not moderating at all.[38] But this apparently exceptional content moderation during the pandemic was only a more exaggerated version of how content moderation works all the time.[39]

What this paradigm shift means for platform governance and its regulation remains undertheorized but is especially important to examine now for two reasons. First, without adapting speech governance to the very different nature of the task being undertaken—systemic balancing instead of individual categorization—platform decisionmaking processes and the rules that govern online speech will continue to be viewed as illegitimate. Because there is no "right" answer to most, if not all, of the questions involved in writing rules for online speech, the rule-formation process is especially important for garnering public acceptance and legitimacy.[40]

---

33. See infra section I.C.1.

34. Evelyn Douek, Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech, 94 Austl. L.J. 41, 50 n.77 (2020) [hereinafter Douek, Nerd Harder].

35. See, e.g., Bickert, Charting a Way Forward, supra note 32; French Sec'y of State for Digit. Affs., Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision: Regulation of Social Networks—Facebook Experiment 13–14 (2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf [https://perma.cc/UAM5-ZP5K] (advocating for public policy that balances punitive and preventative approaches).

36. See infra section I.D.

37. See infra section I.D.

38. See infra section I.D.2.

39. Evelyn Douek, COVID-19 and Social Media Content Moderation, Lawfare (Mar. 25, 2020), https://www.lawfareblog.com/covid-19-and-social-media-content-moderation [https://perma.cc/6MF4-29PQ] [hereinafter Douek, COVID-19 and Social Media Content Moderation].

40. See Ben Bradford, Florian Grisel, Tracey L. Meares, Emily Owens, Baron L. Pineda, Jacob N. Shapiro, Tom R. Tyler & Danieli Evans Peterman, Report of the Facebook Data Transparency Advisory Group 34–39 (2019), https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf [https://perma.cc/3AKC-UUWP] [hereinafter Facebook Data Transparency Advisory Group] ("Facebook could build public trust and legitimacy . . . by following principles of procedural justice in its interactions with users."); Tom R. Tyler, Procedural Justice, Legitimacy, and the Effective

Second, regulators around the world are currently writing laws to change the regulatory landscape for online speech. In the United States in particular, the law that "created the internet"[41]—Section 230 of the Communications Decency Act[42]—is increasingly under siege across the political spectrum, with its reform seemingly imminent.[43] But changing the regulatory environment without a proper understanding of content moderation in practice will make the laws ineffective or, worse, create unintended consequences. Regulators need to understand the inherent characteristics of the systems they seek to reform. Regulation that entrenches one right or interest without acknowledging the empirical realities of how online speech operates and is constantly changing, or that adopts a punitive approach focused on individual cases, will fail to bring the accountability that is currently lacking from platforms without necessarily protecting those harmed by their decisions.[44] This Article therefore offers an account of the role of proportionality and probability in online speech governance and the questions it raises for such governance and its regulation.

This Article concentrates on tech platforms' role as the current primary rulemakers and enforcers of online content regulation, the focus of a rapidly growing literature.[45] This is for two reasons. First, content moderation will always go beyond what governments can constitutionally provide for. The First Amendment would not permit laws requiring removal of content like the Christchurch Massacre livestream,[46] violent

---

Rule of Law, 30 Crime & Just. 283, 284 (2003) (highlighting several studies that suggest "people's willingness to accept the constraints of law . . . is strongly linked to their evaluations of the procedural justice of the police and the courts"); Rory Van Loo, Federal Rules of Platform Procedure, U. Chi. L. Rev. (forthcoming) (manuscript at 28), https://ssrn.com/abstract=3576562 (on file with the *Columbia Law Review*) ("[T]here is strong evidence that the added trust and legitimacy gained from effective dispute resolution systems improves a company's profitability due to better customer retention and increased customer engagement.").

41. Jeff Kosseff, The Twenty-Six Words that Created the Internet 8 (2019).

42. 47 U.S.C. § 230 (2018).

43. See, e.g., Editorial, Section 230 Does Not Need a Revocation. It Needs a Revision., Wash. Post (June 28, 2020), https://www.washingtonpost.com/opinions/trump-and-biden-both-want-to-repeal-this-tech-rule-theyre-both-wrong/2020/06/28/4de6f9fc-b4b1-11ea-a8da-693df3d7674a_story.html (on file with the *Columbia Law Review*) (noting that both President Trump and Joe Biden have called for the repeal of Section 230).

44. See infra section III.B.2.

45. See Hannah Bloch-Wehba, Automation in Moderation, Cornell Int'l L.J. (forthcoming 2020) (manuscript at 4 n.11), https://ssrn.com/abstract=3521619 (on file with the *Columbia Law Review*) [hereinafter Bloch-Wehba, Automation in Moderation]; Van Loo, supra note 40 (manuscript at 3).

46. E.g., Brandenburg v. Ohio, 395 U.S. 444, 447–49 (1969) (holding that a state may forbid speech that advocates violence only if the speech is intended to provoke imminent illegal activity and is likely to do so).

animal crush videos,[47] or graphic pornography,[48] for example, but few would disagree that platforms should have some license to moderate this content to protect their services from becoming unusable. How far this license should extend may be contested, but it is relatively uncontroversial that private actors can restrict more speech than governments. Second, the scale of online content will make private platforms' role as the frontline actors in content moderation an ongoing practical necessity. Governments will not have the resources or technical capacity to take over.

As much as platforms are building bureaucracies and norms in a way that can resemble those of governments,[49] they remain private actors with obvious business interests and are unencumbered by the constraints of public law. The project of online speech governance centers around the question of how to square this triangle[50] of unaccountable private actors exercising enormous power over systemically important public communication while accepting the constitutional and practical limitations of government regulation. This Article's contribution to that task is to describe and give a conceptual framework to the radical changes that have occurred in the actual operation of content moderation in the last half decade alone. Part I begins by describing the categorical and individualistic paradigm of early content moderation—what this Article calls its "posts-as-trumps" era—and how this has given way to an era defined by proportionality and probability in online speech governance. This Article argues that this governance based on systemic balancing is both normatively and pragmatically a better fit for the modern realities of online speech. Descriptively, these principles already shape online speech, whether explicitly acknowledged or not. A case study of platform content moderation during the COVID-19 pandemic illustrates this starkly. Part II turns to the questions that governance based on proportionality and probability raises for decisionmakers and shows that the failure to adequately address these has left current governance arrangements fundamentally unstable and unsatisfying.

Part III turns to the urgent project of addressing these deficiencies. This Article argues that, despite first appearances, systemic balancing in online speech governance need not entail a devaluing or deflation of speech rights. In fact, as a methodological approach, it does not demand

47. United States v. Stevens, 559 U.S. 460, 481–82 (2010) (striking down a federal law that criminalized depictions of animal cruelty under the First Amendment's overbreadth doctrine).

48. Am. Booksellers Ass'n v. Hudnut, 771 F.2d 323, 332–34 (7th Cir. 1985), aff'd 475 U.S. 1001 (1986).

49. See generally Kate Klonick, The New Governors: The People, Rules, and Processes Governing Online Speech, 131 Harv. L. Rev. 1598 (2018) (discussing the ways in which internet platforms have developed detailed systems for governing online speech that are rooted in the American legal system).

50. Jack M. Balkin, Free Speech Is a Triangle, 118 Colum. L. Rev. 2011, 2012 (2018); Robert Gorwa, The Platform Governance Triangle: Conceptualising the Informal Regulation of Online Content, Internet Pol'y Rev., June 2019, at 1, 2.

any particular substantive results and could result in more speech-protective rules. The critical point is that recognizing the role of systemic balancing orients debates around the right questions. This Article therefore turns to what these questions are for both platforms and regulators, and discusses their impact on what content moderation should look like in a post-pandemic world.

Online speech governance is a wicked problem with unenviable and perhaps impossible trade-offs. There is no end-state of content moderation with stable rules or regulatory forms; it will always be a matter of contestation, iteration, and technological evolution. That said, this is an unusual period of disruption and experimentation, as the prevailing forms of internet governance have become inadequate and new systems are emerging to replace them. Understanding what tasks these institutions must be designed to fulfill is the first step to evaluating, improving, and regulating them.

I. THE OLD AND NEW PRECEPTS OF ONLINE SPEECH GOVERNANCE

Early online speech governance was animated by what this Article calls a "posts-as-trumps" ethos.[51] This reflected the First Amendment's categorical and individualistic approach to speech adjudication and its conception of "freedom of speech as a classic trump."[52] The starting point was that "the [posts] must flow."[53] The approach was never one of free speech absolutism. But, in general, freedom of expression was exalted,[54] and platforms presumptively allowed "users to post what they wanted."[55] Or, in the now-infamous words of a top Facebook executive, "The ugly truth is that we believe in connecting people so deeply that anything that allows us to connect more people more often is *de facto* good."[56] For the first decade or so, "online intermediaries were avowedly laissez faire about user-generated content."[57] Zittrain has called this "The Rights Era"

51. This coinage follows Ronald Dworkin's famous conception of rights as trumps. Ronald Dworkin, Taking Rights Seriously, at xi, 192 (1977); see also Jamal Greene, Foreword: Rights as Trumps?, 132 Harv. L. Rev. 28, 36 (2018).

52. Greene, supra note 51, at 36.

53. Marvin Ammori, The "New" *New York Times*: Free Speech Lawyering in the Age of Google and Twitter, 127 Harv. L. Rev. 2259, 2260 (2014) (cleaned up) (internal quotation marks omitted) (citing Biz Stone, The Tweets Must Flow, Twitter Blog (Jan. 28, 2011), https://blog.twitter.com/2011/tweets-must-flow [https://perma.cc/BTB5-NJ3P]).

54. Olivier Sylvain, Recovering Tech's Humanity, 119 Colum. L. Rev. Forum 252, 255–56 (2019) ("For the first decade or so after the commercial deployment of the internet . . . online intermediaries were avowedly laissez faire about user-generated content . . . . It was an exciting time for . . . free speech advocates.").

55. Jillian C. York & Ethan Zuckerman, Moderating the Public Sphere, *in* Human Rights in the Age of Platforms 137, 143 (Rikke Frank Jørgensen ed., 2019).

56. Sheera Frenkel & Nellie Bowles, Facebook Employees in an Uproar over Executive's Leaked Memo, N.Y. Times (Mar. 30, 2018), https://www.nytimes.com/2018/03/30/technology/facebook-leaked-memo.html (on file with the *Columbia Law Review*).

57. Sylvain, supra note 54, at 255.

of internet governance, dominated by a "classic libertarian ethos" of preserving individual affordances in speech with which intermediaries should interfere to a limited extent.[58]

This Part starts by describing the reasons for that early approach. This story is fairly well told, and so it is rehearsed here only briefly before showing how this framework became strained along two dimensions. First, there were increased regulatory and social demands for platforms to take account of interests other than the individual speech right which led to the rise of a proportionality approach to content moderation. Second, the unfathomable scale of the major tech platforms has made an individualistic conception of speech adjudication untenable, leading to a systemic and probabilistic approach to speech governance. This Part describes the nature of these two shifts in turn and defends the need to center them in online rights adjudication as the only realistic future for platform governance. Finally, this Part concludes with a case study of content moderation during the COVID-19 pandemic as epitomizing the rise of proportionality and probability in content moderation.

A. *Posts-As-Trumps*

It is now familiar to observe that early online speech governance was highly influenced by the First Amendment tradition to which lawyers at the major platforms were accustomed.[59] As private companies, platforms are not subject to First Amendment constraints,[60] and commercial interests no doubt guided (and continue to guide) formulation of platforms' rules for what they permit on their sites.[61] Nevertheless, the First Amendment influence on platforms was profound, for three reasons: (1) Broad legal immunities allowed platforms to take a hands-off approach; (2) U.S. norms were legally and culturally familiar; and (3) most importantly for present purposes, these norms methodologically follow a categorical approach which allows decisionmakers to profess humility and suggest they are not making value judgments about any particular speech.

1. *The Role of Formal Law.* — The early days of the internet are sometimes referred to as the "laissez-faire" era of internet regulation,[62] but this should not be taken to mean that formal law was inconsequential. Law's role during this time was passive, but it deeply shaped speech

---

58. Zittrain, Three Eras, supra note 21, at 1.

59. See Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media 12 (2018) [hereinafter Gillespie, Custodians of the Internet]; David Kaye, Speech Police: The Global Struggle to Govern the Internet 47 (2019) [hereinafter Kaye, Speech Police]; Klonick, supra note 49, at 1621.

60. Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019).

61. Gillespie, Custodians of the Internet, supra note 59, at 5; Klonick, supra note 49, at 1627.

62. See, e.g., David Kaye, Foreword, *in* Human Rights in the Age of Platforms, supra note 55, at xi, xii [hereinafter Kaye, Human Rights].

governance.[63] The internet was never unregulated—instead, laws created immunities and safe harbors for internet intermediaries like platforms in order to facilitate the free flow of expression and foster innovation.[64] That is: Law deliberately created a space where posts could flow and trump other interests.[65]

Platforms embraced the opportunity law gave them to take a light-touch approach to moderation for a simple reason: Platforms run on speech. To put it simply, "The core business functions of Twitter, YouTube, and other platforms turn on expression . . . . The lawyers working for these companies have business reasons for supporting free expression."[66] Or, as Professor Noah Feldman put it to Facebook CEO Mark Zuckerberg, "No voice, no Facebook, right?"[67]

2. *Cultural and Legal Familiarity.* — Tech companies are famously monocultural, and rules for their global services were set in Silicon Valley by lawyers "acculturated in American free speech norms."[68] As an early Google and Twitter lawyer put it, their products were built from a "particular set of perspectives (and that's what the [First Amendment] norms are probably part of) that was nowhere near diverse enough given the eventual reach and importance of our products."[69] Accordingly, these lawyers held a distinctively American conception of free speech. The allure still holds: When Facebook CEO Mark Zuckerberg "took a stand" for voice and free expression at Georgetown University in 2019, it was to the First Amendment that he turned in explaining the importance of protecting speech, noting that he was "proud that our values at Facebook are inspired by the American tradition, which is more supportive of free expression than anywhere else."[70] The stickiness of this cultural familiarity, even as

63. See Kosseff, supra note 41, at 3 ("Section 230 created the legal and social framework for the Internet we know today . . . .").

64. See Anupam Chander, How Law Made Silicon Valley, 63 Emory L.J. 639, 648–50 (2014).

65. Hannah Bloch-Wehba, Global Platform Governance: Private Power in the Shadow of the State, 72 SMU L. Rev. 27, 33–39 (2019) (reviewing the self-governance ethos that dominated the early internet and the ways it was enabled by regulation); Sylvain, supra note 54, at 253–54 ("There has never been the chance to see what even modest run-of-the-mill judicial adjudication of content moderation decisions looks like since Congress enacted section 230 over twenty years ago.").

66. Ammori, supra note 53, at 2260.

67. A Conversation with Mark Zuckerberg, Jenny Martinez and Noah Feldman, Facebook (June 27, 2019), https://about.fb.com/news/2019/06/mark-challenge-jenny-martinez-noah-feldman [https://perma.cc/ZX3V-BUDV].

68. Klonick, supra note 49, at 1621.

69. Alexander Macgillivray, First Amendment and Earlyish Content Moderation, Bricoleur (May 7, 2020), http://www.bricoleur.org/2020/05/first-amendment-and-early ish-content.html [https://perma.cc/L63A-ARN2].

70. Mark Zuckerberg Stands for Voice and Free Expression, Facebook: Newsroom (Oct. 17, 2019), https://about.fb.com/news/2019/10/mark-zuckerberg-stands-for-voice-and-free-expression [https://perma.cc/DFF3-6R7M].

these companies cannot really be described as anything but global, illustrates its profound influence.

3. *Methodological Exceptionalism and Professed Humility.* — So far, this is all a familiar story. But one aspect of the First Amendment tradition's influence on early online speech governance has gone underappreciated: U.S. free speech jurisprudence is not only substantively the most speech-protective in the world—it is also *methodologically* exceptional.[71] Highlighting this methodological difference—which is best summarized as approaching the task as one of categorization rather than involving overt weighing and balancing of interests—is important because it underscores the very different role platforms are required to perform in the post-"posts-as-trumps" era.

U.S. free speech adjudication is marked by a two-step categorical approach.[72] First, the decisionmaker asks whether or not the speech fits into a category covered by the First Amendment. Second, a series of fairly outcome-determinative rules are applied based on this categorization. The hard work under this framework is therefore done at the initial step of defining the scope of rights through the use of categories.[73] As noted earlier, under this framework speech rights are classic "trumps,"[74] following Ronald Dworkin's argument that to subject them to balancing against other interests is to deny them altogether.[75] The general stance of U.S. First Amendment law is that it is "startling and dangerous" to protect speech through a free-floating test based on ad hoc balancing of relative social costs and benefits, and therefore restrictions on speech should be limited to a set of narrowly confined categories.[76] Famously, under this approach, "the best test of truth is the power of the thought to get itself accepted in the competition of the market."[77]

---

71. See Frederick Schauer, The Exceptional First Amendment, *in* American Exceptionalism and Human Rights 29, 30–32 (Michael Ignatieff ed., 2005) [hereinafter Schauer, Exceptional First Amendment].

72. See Adrienne Stone, The Comparative Constitutional Law of Freedom of Expression, *in* Comparative Constitutional Law 406, 410 (Tom Ginsburg & Rosalind Dixon eds., 2011) [hereinafter Stone, Comparative Freedom of Expression] ("[T]he American doctrine of the First Amendment is characterized by a 'conceptual' or 'categorical' approach, according to which freedom of expression law is dominated by relatively inflexible rules, each with application to a defined category of circumstances.").

73. Aharon Barak, Proportionality: Constitutional Rights and Their Limitations 504 (2012) [hereinafter Barak, Constitutional Rights and Their Limitations] ("The main legal issue thus becomes the identification of the proper category, and then the application of the factual framework to that proper, pre-determined legal category. Once a category has been chosen, the accompanying set of legal rules will automatically apply.").

74. Greene, supra note 51, at 36.

75. See Dworkin, supra note 51, at xi, 192.

76. United States v. Alvarez, 567 U.S. 709, 717 (2012); United States v. Stevens, 559 U.S. 460, 470 (2010); see also R.A.V. v. City of St. Paul, 505 U.S. 377, 383 (1992) ("[A] limited categorical approach has remained an important part of our First Amendment jurisprudence.").

77. Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

"Posts-as-trumps" was not free speech absolutism, but neither is First Amendment jurisprudence. Indeed, "No responsible approach to the problem [of free speech] can be oblivious to the dangers certain types of expression pose."[78] From the start, content moderation was "fueled by this contradiction: an ambivalence about intervening, and a fear of not intervening."[79] All major platforms have always restricted some speech that is legal under the First Amendment. But the exceptionally high value on freedom of expression embodied in American constitutional law[80] was reflected in early platform rules which erred on the side of limited, categorical exceptions. Facebook's first set of Community Standards, for example, "generally adhered to John Stuart Mill's seminal principle that speech should be banned only if used to stoke violence against others."[81] Twitter's first rules were a mere 568 words and gradually expanded to give limited categorical exceptions to the free speech ethos, often based around account behavior targeting spam,[82] in a way that might be analogized to "time, place and manner" restrictions.[83] YouTube's first content moderators were given a one-page list of instructions setting out a list of categories of content they should remove.[84]

Even as platforms have developed more elaborate rules, elements of this categorical approach remain. The starkest illustration of this is the refusal to remove false content.[85] Platforms regularly protest they should

78. John Hart Ely, Democracy and Distrust: A Theory of Judicial Review 110 (1980).

79. Gillespie, Custodians of the Internet, supra note 59, at 50.

80. Stone, Comparative Freedom of Expression, supra note 72, at 411.

81. Simon van Zuylen-Wood, "Men Are Scum": Inside Facebook's War on Hate Speech, Vanity Fair (Feb. 26, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech (on file with the *Columbia Law Review*).

82. See Sarah Jeong, The History of Twitter's Rules, Vice (Jan. 14, 2016), https://www.vice.com/en_us/article/z43xw3/the-history-of-twitters-rules [https://perma.cc/V4UG-AVHS].

83. Ward v. Rock Against Racism, 491 U.S. 781, 791 (1989).

84. Catherine Buni & Soraya Chemaly, The Secret Rules of the Internet, Verge, https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech (on file with the *Columbia Law Review*) (last visited Oct. 23, 2020).

85. Although platforms have partnerships with third-party fact-checkers, these result in warning labels and context being attached to content that has been fact-checked, as well as reduced circulation, rather than removal of that content. See, e.g., Daniel Funke, YouTube Is Now Surfacing Fact Checks in Search. Here's How It Works., Poynter (Mar. 8, 2019), https://www.poynter.org/fact-checking/2019/youtube-is-now-surfacing-fact-checks-in-search-heres-how-it-works [https://perma.cc/5KDL-TS23]; Tessa Lyons, Hard Questions: How Is Facebook's Fact-Checking Program Working?, Facebook: Newsroom (June 14, 2018), https://about.fb.com/news/2018/06/hard-questions-fact-checking [https://perma.cc/Z44Q-C26C] [hereinafter Lyons, Hard Questions].

not be "arbiters of truth."[86] This has, until recently,[87] manifested as a categorical refusal to remove content simply because it is untrue—regardless of context, how potentially harmful the content, or how readily disprovable the lie.[88] A well-known example is Facebook's staunch refusal—only recently revised[89]—to remove Holocaust denial, because "there are things that different people get wrong."[90] This tracks the First Amendment's position that falsehoods are protected because "some false statements are inevitable if there is to be an open and vigorous expression of views."[91]

Another example of categorical thinking is Facebook's treatment of adult nudity. Facebook has long had a largely no-tolerance approach to nudity. The company has caused outrage by taking down everything from paintings,[92] to Pulitzer Prize–winning photos,[93] to breastfeeding mothers.[94] The contestation around nudity on Facebook has been about definition of the category. Activists have had some success in getting Facebook to create limited exceptions to the general ban,[95] but they remain narrow. The

---

86. See, e.g., Callum Borchers, Twitter Executive on Fake News: 'We Are Not the Arbiters of Truth', Wash. Post (Feb. 8, 2018), https://www.washingtonpost.com/news/the-fix/wp/2018/02/08/twitter-executive-on-fake-news-we-are-not-the-arbiters-of-truth (on file with the *Columbia Law Review*); Supraja Srinivasan, We Don't Want to Be Arbiters of Truth: YouTube CBO Robert Kyncl, Econ. Times (Mar. 24, 2018), https://tech.economictimes.indiatimes.com/news/internet/we-dont-want-to-be-arbiters-of-truth-youtube-cbo-robert-kyncl/63438805 [https://perma.cc/E86J-RXYU]; Mark Zuckerberg, Facebook Status Update, Facebook (Nov. 18, 2016), https://www.facebook.com/zuck/posts/10103269806149061 [https://perma.cc/AQ2F-S6EM] [hereinafter Zuckerberg, Status Update].

87. See infra section I.D.

88. But, as infra section I.D discusses, the pandemic has forced platforms to touch even this third rail of content moderation.

89. Monika Bickert, Removing Holocaust Denial Content, Facebook Newsroom (Oct. 12, 2020), https://about.fb.com/news/2020/10/removing-holocaust-denial-content [https://perma.cc/8M6U-6XPK] [hereinafter Bickert, Removing Holocaust Denial Content].

90. Kara Swisher, Zuckerberg: The Recode Interview, Vox, https://www.vox.com/2018/7/18/17575156/mark-zuckerberg-interview-facebook-recode-kara-swisher (on file with the *Columbia Law Review*) (last updated Oct. 8, 2018).

91. United States v. Alvarez, 567 U.S. 709, 718 (2012).

92. Amar Toor, 19th Century Vagina Sparks French Lawsuit Against Facebook, Verge (Mar. 6, 2015), https://www.theverge.com/2015/3/6/8160721/facebook-censorship-vagina-painting-france-lawsuit (on file with the *Columbia Law Review*).

93. James Vincent, Zuckerberg Criticized over Censorship After Facebook Deletes 'Napalm Girl' Photo, Verge (Sept. 9, 2016), https://www.theverge.com/2016/9/9/12859686/facebook-censorship-napalm-girl-aftenposten (on file with the *Columbia Law Review*).

94. Soraya Chemaly, #FreeTheNipple: Facebook Changes Breastfeeding Mothers Photo Policy, HuffPost (June 9, 2014), https://www.huffpost.com/entry/freethenipple-facebook-changes_b_5473467 [https://perma.cc/CQ4K-HVFH] (last updated Dec. 6, 2017); Julia Jacobs, Will Instagram Ever 'Free the Nipple'?, N.Y. Times (Nov. 22, 2019), https://www.nytimes.com/2019/11/22/arts/design/instagram-free-the-nipple.html (on file with the *Columbia Law Review*).

95. Chemaly, supra note 94.

classification of a post as nudity is generally outcome-determinative: It comes down.

A categorical approach therefore does not always have to be more speech-protective. Facebook and Twitter's different rules on political advertising demonstrate this—neither wants to be drawn into becoming a referee of highly charged and contested political debates, both because it can be practically difficult and because it could alienate large segments of their user base.[96] Put another way, "Republicans buy shoes from Facebook ads too."[97] In order to hew to this notion that they do not slant the playing field, both platforms have adopted categorical rules, but at different extremes. Facebook does not let politicians' ads be fact-checked:[98] Political ads are categorically protected. Twitter instead categorically bans political ads altogether.[99]

This methodological approach not only has substantive consequences, but also casts the decisionmaker in a different role. Under this frame, the decisionmaker asserts a kind of neutrality, professing to be a mere "taxonomist[]."[100] A categorical approach "induces our identification of rights to track the categories judges are able to access, articulate, and delimit rather than the moral, political, or even constitutional justice the rights mean to promote."[101] Indeed, First Amendment jurisprudence reflects the Court's "'fear' of making obvious judgment calls on matters of degree."[102] The Deciders[103] of early online speech governance found themselves thrust, often unexpectedly, into a position of presiding over the speech of millions and wanted to avoid contradicting baselines drawn by legal systems or wading into hard moral and political fights.[104] Adopting

---

96. See supra note 85.

97. Evelyn Douek, Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility, 21 N.C. J.L. & Tech. 1, 26 (2019).

98. Nick Clegg, Facebook, Elections and Political Speech, Facebook: Newsroom (Sept. 24, 2019), https://newsroom.fb.com/news/2019/09/elections-and-political-speech [https://perma.cc/A2GX-XM9S].

99. Kate Conger, Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says, N.Y. Times (Oct. 30, 2019), https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html (on file with the *Columbia Law Review*).

100. Sullivan, supra note 17, at 293.

101. Greene, supra note 51, at 32–33.

102. Jackson, supra note 15, at 3149 (quoting Mark Tushnet, The First Amendment and Political Risk, 4 J. Legal Analysis 103, 114 (2012)).

103. Jeffrey Rosen, Lecture, The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google, 80 Fordham L. Rev. 1525 (2012).

104. See Macgillivray, supra note 69 ("[W]e felt that we were relatively new to analyzing this stuff but that . . . the Courts had more experience drawing those lines . . . ."); Andrew Marantz, The Dark Side of Techno-Utopianism, New Yorker (Sept. 23, 2019), https://www.newyorker.com/magazine/2019/09/30/the-dark-side-of-techno-utopianism (on file with the *Columbia Law Review*) ("I remember thinking, [p]eople in government, on the Supreme Court, are way smarter than me . . . . So, if something's not illegal to say under U.S. law, why should I make it illegal to say on Reddit?" (internal quotation marks omitted) (quoting Steve Huffman, cofounder and Chief Executive Officer of Reddit)).

a categorical stance therefore allowed early platform lawyers to profess that they were not "making judgment calls about the value of particular speech."[105]

In sum, "posts-as-trumps" generally held that speech should be as free as possible, and exceptions narrow and rule-based.[106] This approach minimizes attention to particular context: Decisionmaking is primarily a matter of deferring to precedential categorizations in the rules. Platforms adopted a position of professed neutrality, expressed as humility: "Who are we to know?"[107] As one early platform lawyer put it, this stance was "both philosophical (who are we to judge?) and pragmatic (no automated system could accurately screen [everything] uploaded each year at scale)."[108] Both these philosophical and pragmatic assumptions have come under strain, as shown in turn in the next two sections.

B. *The Rise of Proportionality*

1. *Reasons Behind the Rise.* — Platforms' professed humility has become untenable. Humility sits uneasily with unilateral authority, and platforms are law-writers, -enforcers, and -reviewers, all in one. There is no check, balance, or democratic accountability for the rules they write and administer. As several platforms have increasingly claimed to be, and have come to be seen as, public squares of systemic importance,[109] their humility from a position of such power has rung especially hollow with regulators concerned about the effects of their services and with the public at large.[110] The "techlash" of the past few years is a general reflection of this

---

105. Ammori, supra note 53, at 2276.

106. See supra notes 56–63 and accompanying text.

107. Ammori, supra note 53, at 2276–77.

108. Ava Kofman, Francis Tseng & Moira Weigel, The Hate Store: Amazon's Self-Publishing Arm Is a Haven for White Supremacists, ProPublica (Apr. 7, 2020), https://www.propublica.org/article/the-hate-store-amazons-self-publishing-arm-is-a-haven-for-white-supremacists [https://perma.cc/KZ48-6TW2].

109. U.S. H. Comm. on Energy & Com., Testimony of Jack Dorsey, Chief Executive Officer, Twitter, Inc. 1 (2018), https://docs.house.gov/meetings/IF/IF00/20180905/108642/HHRG-115-IF00-Wstate-DorseyJ-20180905.pdf [https://perma.cc/UJ89-935S] ("Twitter is used as a global town square, where people from around the world come together in an open and free exchange of ideas."); Mark Zuckerberg, A Privacy-Focused Vision for Social Networking, Facebook, https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634 [https://perma.cc/2FS4-XS8H] (last updated Mar. 6, 2019) ("Over the last 15 years, Facebook and Instagram have helped people connect with friends, communities, and interests in the digital equivalent of a town square.").

110. See, e.g., John Laloggia, U.S. Public Has Little Confidence in Social Media Companies to Determine Offensive Content, Pew Rsch. Ctr. (July 11, 2019), https://www.pewresearch.org/fact-tank/2019/07/11/u-s-public-has-little-confidence-in-social-media-companies-to-determine-offensive-content [https://perma.cc/8QJU-4SNS] ("A sizable majority of U.S. adults (66%) say social media companies have a responsibility to remove offensive content from their platforms . . . .").

sentiment,[111] but the shift to a proportionality-based approach to content moderation has four more specific foundations: (a) exposure of the "myth of platform neutrality;"[112] (b) a cultural shift in attitudes toward online speech; (c) the incoherence in platforms' proliferating categories; and (d) the need for platforms that were increasingly global to acknowledge the globally dominant approach to rights adjudication: proportionality.

a. *The Myth of Platform Neutrality.* — One lesson platforms could have learned from First Amendment jurisprudence is that the categorical approach does not allow the decisionmaker to remain neutral for long. "[C]ategories are not natural objects,"[113] and the process of categorization leaves ample space for value judgments. As much as they profess to, even under a categorical approach decisionmakers "cannot . . . escape judgment."[114] This has been a sustained line of attack on First Amendment doctrine,[115] and it only applies more forcefully to platforms that are far from impartial about how content is presented to users.

Speech on platforms is a complex interaction of user interests, platform affordances, and algorithmic choices. As Jameel Jaffer wrote, "Facebook's users interact and speak to one another in an environment shaped by Facebook's interface, algorithms, and policies. What gets said is up to individual users, but it's Facebook that determines which speech is amplified and which is suppressed."[116] If the "marketplace of ideas" analogy was ever more than an evocative oversimplification,[117] it surely does not apply to platform ecosystems that optimize for engagement rather than truth.[118] Optimizing for engagement elevates "glimmers of

111. Eve Smith, A Memo to Big Tech: The Techlash Against Amazon, Facebook and Google—And What They Can Do, Economist (Jan. 20, 2018), https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do (on file with the *Columbia Law Review*).

112. Anupam Chander & Vivek Krishnamurthy, The Myth of Platform Neutrality, 2 Geo. L. Tech. Rev. 400, 400 (2018).

113. Mark Tushnet, The First Amendment and Political Risk, 4 J. Legal Analysis 103, 116 (2012).

114. Jackson, supra note 15, at 3192.

115. See, e.g., Genevieve Lakier, The Invention of Low-Value Speech, 128 Harv. L. Rev. 2166, 2173–77, 2233 (2015) (recounting the "persistent" controversy around the categories of low-value speech, and concluding that "courts cannot avoid the difficult task of judging the constitutional value of novel categories of speech").

116. Jameel Jaffer, Facebook and Free Speech Are Different Things, Knight First Amend. Inst. (Oct. 24, 2019), https://knightcolumbia.org/content/facebook-and-free-speech-are-different-things [https://perma.cc/PRM3-Y5R4].

117. See Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting) ("[M]en . . . may come to believe . . . that the ultimate good desired is better reached by free trade in ideas—that the best test of truth is the power of the thought to get itself accepted in the competition of the market . . . .").

118. See, e.g., Cass R. Sunstein, #Republic: Divided Democracy in the Age of Social Media 3–6 (2017) ("[T]he architecture of control [by social media platforms] has a serious downside, raising fundamental questions about freedom, democracy, and self-government."); Siva Vaidhyanathan, Antisocial Media: How Facebook Disconnects Us and

novelty, messages of affirmation and belonging, and messages of outrage toward perceived enemies" regardless of veracity or social value.[119] There is nothing natural or inevitable about these priorities in platform design and the way they direct users' attention.[120] Such intentional distortions make platforms' asserted "humility" seem, at best, an abdication of responsibility and, at worst, an outright lie.

b. *Cultural and Regulatory Shifts.* — There has also been a broader cultural reevaluation, and "the public's sense of what platforms ought to be held accountable for has shifted tectonically."[121] As cyber–civil rights pioneer Danielle Citron has observed, "Much has changed in the past ten years," and there has been a marked societal turn away from the thin conceptions of online speech to allow for greater recognition of other interests.[122] A categorical approach to rights raises "the danger of harmful over-enforcement of rights deemed fundamental."[123] This danger has been salient in recent years, as the vast quantities of toxic speech online have been more visible than ever.[124] Platforms have found themselves

---

Undermines Democracy 4 (2018) ("[T]he idealistic vision of people sharing more information with ever more people has not improved nations or global culture, enhanced mutual understanding, or strengthened democratic movements."); Zeynep Tufekci, It's the (Democracy-Poisoning) Golden Age of Free Speech, WIRED (Jan. 16, 2018), https://www.wired.com/story/free-speech-issue-tech-turmoil-new-censorship [https://perma.cc/S232-6KWS] [hereinafter Tufekci, Golden Age of Free Speech] ("Many more of the most noble old ideas about free speech simply don't compute in the age of social media. John Stuart Mill's notion that a 'marketplace of ideas' will elevate the truth is flatly belied by the virality of fake news.").

119. Tufekci, Golden Age of Free Speech, supra note 118.

120. Id.

121. Bowers & Zittrain, supra note 19, at 2.

122. Danielle Keats Citron, Restricting Speech to Protect It, *in* Free Speech in the Digital Age 122, 122 (Susan J. Brison & Katherine Gelber eds., 2019) [hereinafter Citron, Restricting Speech]; see also Gillespie, Custodians of the Internet, supra note 59, at 36 ("[A] slow reconsideration of platform responsibility has been spurred by categories of content particularly abhorrent to users and governments.").

123. Mattias Kumm, Constitutional Rights as Principles: On the Structure and Domain of Constitutional Justice, 2 Int'l J. Const. L. 574, 595 (2004) [hereinafter Kumm, Constitutional Rights as Principles].

124. See, e.g., Andrew Marantz, Opinion, Free Speech Is Killing Us, N.Y. Times (Oct. 4, 2019), https://www.nytimes.com/2019/10/04/opinion/sunday/free-speech-social-media-violence.html (on file with the *Columbia Law Review*); Richard Stengel, Opinion, Why America Needs a Hate Speech Law, Wash. Post (Oct. 29, 2019), https://www.washingtonpost.com/opinions/2019/10/29/why-america-needs-hate-speech-law (on file with the *Columbia Law Review*); Tufekci, Golden Age of Free Speech, supra note 118; Tim Wu, Is the First Amendment Obsolete?, Knight First Amend. Inst. (Sept. 1, 2017), https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete [https://perma.cc/5DK2-R8PK].

implicated in controversies as diverse as genocide,[125] election interference,[126] widespread harassment and abuse,[127] and terrorist attacks,[128] to name but a few examples. As a result, there have been growing calls for them to mitigate harm from their services, and "tech industry leaders have finally begun to take some online abuses seriously and to reckon with the detrimental impact they have on democracy, autonomy, and truth. More and more online platforms are confronting the reality that the best answer to bad speech is not, in fact, more speech."[129] As Twitter's general counsel acknowledged, it was necessary to *balance* "welcoming diverse perspectives while protecting our users" because "[f]reedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up."[130]

Even regarding hate speech—the steadfast protection of which is almost synonymous with First Amendment exceptionalism[131]—platforms have dramatically changed their approach. One stark example is the sudden, seemingly arbitrarily timed ban of high-profile, far-right conspiracy theorist Alex Jones from several major platforms.[132] Since then,

125. See, e.g., Evelyn Douek, Facebook's Role in the Genocide in Myanmar: New Reporting Complicates the Narrative, Lawfare (Oct. 22, 2018), https://www.lawfareblog.com/facebooks-role-genocide-myanmar-new-reporting-complicates-narrative [https://perma.cc/UB39-UM35]; Evelyn Douek, Why Were Members of Congress Asking Mark Zuckerberg About Myanmar? A Primer., Lawfare (Apr. 26, 2018), https://www.lawfareblog.com/why-were-members-congress-asking-mark-zuckerberg-about-myanmar-primer [https://perma.cc/F3Y4-GLMR].

126. See, e.g., Renee DiResta, Kris Shaffer, Becky Ruppel, David Sullivan, Robert Matney, Ryan Fox, Jonathan Albright & Ben Johnson, The Tactics & Tropes of the Internet Research Agency 100–01 (2019), https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf [https://perma.cc/9TSC-8TUX] (noting the role of social media platforms in voter disinformation campaigns).

127. Danielle Keats Citron & Mary Anne Franks, The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform, at 9 (Bos. Univ., Pub. L. & Legal Theory Working Paper No. 20-8, 2020), https://ssrn.com/abstract=3532691 (on file with the *Columbia Law Review*) ("Section 230 has subsidized platforms whose business is online abuse.").

128. See, e.g., Charlie Warzel, Opinion, The New Zealand Massacre Was Made to Go Viral, N.Y. Times (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/opinion/new-zealand-shooting.html (on file with the *Columbia Law Review*).

129. Mary Anne Franks, The Free Speech Black Hole: Can the Internet Escape the Gravitational Pull of the First Amendment?, Knight First Amend. Inst. (Aug. 21, 2019), https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment [https://perma.cc/7RBL-SC3W] [hereinafter Franks, The Free Speech Black Hole].

130. Vijaya Gadde, Twitter Executive: Here's How We're Trying to Stop Abuse While Preserving Free Speech, Wash. Post (Apr. 16, 2015), https://www.washingtonpost.com/posteverything/wp/2015/04/16/twitter-executive-heres-how-were-trying-to-stop-abuse-while-preserving-free-speech (on file with the *Columbia Law Review*).

131. See Schauer, Exceptional First Amendment, supra note 71, at 32–38.

132. See Jane Coaston, YouTube, Facebook, and Apple's Ban on Alex Jones, Explained, Vox (Aug. 6, 2018), https://www.vox.com/2018/8/6/17655658/alex-jones-facebook-you

Facebook,[133] YouTube,[134] and Twitter[135] have all released stricter hate speech policies. A *New York Times* opinion piece even declared it "relatively uncontroversial" that these platforms should remove hate speech.[136] What caused this marked change in public discourse is, of course, a much deeper question that will no doubt be probed elsewhere. Suffice it to say that faced with high-profile incidents of rising violence and threats of violence tied to online hate speech, there was a "tidal shift" of opinion toward taking such speech offline entirely.[137]

Perhaps the clearest symbol of the shift from "posts-as-trumps" to proportionality is Facebook's decision to ban Holocaust denial.[138] There is no more emblematic case of the American approach to freedom of expression than the *Skokie* case in the late 1970s,[139] which held that it was inconsistent with the First Amendment to deny the right of neo-Nazis to march in a suburb of Chicago with a large Jewish population.[140] The case is iconic in U.S. free speech lore and is considered one of the "truly great victories for the First Amendment."[141] Even so, Facebook's steadfast refusal to ban such content was, for a long time, its "most controversial policy,"[142]

---

tube-conspiracy-theories (on file with the *Columbia Law Review*); Franks, The Free Speech Black Hole, supra note 129. What is striking about the Jones incident is that the content that led almost every major online platform to take action against him in the middle of 2018 was not substantially different than content he had been posting for years. Public pressure, however, increased, and once Apple took the step of banning Jones and his content entirely, there was a domino effect across a wide range of platforms. Coaston, supra.

133. Standing Against Hate, Facebook: Newsroom (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate [https://perma.cc/PE8X-2645].

134. Our Ongoing Work to Tackle Hate, YouTube Off. Blog (June 5, 2019), https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html [https://perma.cc/375S-J79S].

135. Updating Our Rules Against Hateful Conduct, Twitter Blog (July 9, 2019), https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html [https://perma.cc/6BMP-3E76] (last updated Mar. 5, 2020).

136. Brittan Heller, Opinion, Is This Frog a Hate Symbol or Not?, N.Y. Times (Dec. 24, 2019), https://www.nytimes.com/2019/12/24/opinion/pepe-frog-hate-speech.html (on file with the *Columbia Law Review*).

137. See, e.g., Quinta Jurecic, Gab Vanishes, and the Internet Shrugs, Lawfare (Oct. 29, 2018), https://www.lawfareblog.com/gab-vanishes-and-internet-shrugs [https://perma.cc/B33G-6PBP] (describing this "change of heart" and the possible causes).

138. See Bickert, Removing Holocaust Denial Content, supra note 89.

139. See Michel Rosenfeld, Hate Speech in Constitutional Jurisprudence: A Comparative Analysis, 24 Cardozo L. Rev. 1523, 1537 (2003) [hereinafter Rosenfeld, Hate Speech in Constitutional Jurisprudence] ("If one case has come to symbolize the contemporary political and constitutional response to hate speech in the United States, it is the Skokie case . . . .").

140. See Smith v. Collin, 436 U.S. 953, 953 (1978) (denying certiorari); Nat'l Socialist Party of Am. v. Village of Skokie, 432 U.S. 43, 43–44 (1977).

141. Geoffrey R. Stone, Remembering the Nazis in Skokie, HuffPost (May 20, 2009), https://www.huffpost.com/entry/remembering-the-nazis-in_b_188739 [https://perma.cc/5TXZ-UEXN] (last updated May 25, 2011).

142. Casey Newton, How Real-World Violence Led Facebook to Overturn Its Most Controversial Policy, Verge (Oct. 14, 2020), https://www.theverge.com/2020/10/14/

and a proxy for its free speech bona fides. No doubt as a result of continued public pressure, and evidence of rising online hate speech and anti-Semitism globally,[143] Facebook reversed its stance. In announcing the ban, Mark Zuckerberg stated, "Drawing the right lines between what is and isn't acceptable speech isn't straightforward, but with the current state of the world, I believe this is the right *balance*."[144] This episode perfectly encapsulates the general trend from First Amendment exceptionalism to a recognition of the need to *balance* various interests proportionally—in the manner other legal systems do[145]—given the evidence of harm resulting from online speech.

The ascendency of this approach, and evidence that it is here to stay, is reflected in the watershed decision of almost all major platforms to cut ties with President Donald Trump or his campaign following his incitement of insurrection at the Capitol on January 6, 2021.[146] Once unthinkable, and indeed a path that companies strenuously resisted calls to adopt for years, platforms exercised awesome power in deplatforming the leader of the free world, explaining in tortured blog posts how, on balance, the risk of harm in allowing the President to keep speaking freely through their services was just too high.[147] Cast in formalistic terms, their decisions came after immense social pressure.

Formal law of course plays a part in this story too, with this cultural shift beginning to be reflected in legislation. The U.K. Government's White Paper on Online Harms proposes a regulator who will adopt "the principle of proportionality" as a "key element" of their approach.[148] In *Google Spain SL v. Agencia Española de Protección de Datos*,[149] the Court of Justice of the European Union tasked Google with the job of balancing the right to privacy and the right to free expression in "right to be forgotten"

21516088/facebook-holocaust-deniers-policy-qanon-anti-semitism (on file with the *Columbia Law Review*).

143. See Bickert, Removing Holocaust Denial Content, supra note 89.

144. Mark Zuckerberg, Facebook (Oct. 12, 2020), https://www.facebook.com/4/posts/10112455086578451 [https://perma.cc/5K2L-NS2T] (emphasis added).

145. See, e.g., Bundesverfassungsgericht [BVerfG] [Federal Constitutional Court] Apr. 13, 1994, 90 Entscheidungen des Bundesverfassungsgerichts [BVerfGE] 241 (Ger.) (Holocaust denial case).

146. Evelyn Douek, Trump Is Banned. Who Is Next?, Atlantic (Jan. 9, 2021), https://www.theatlantic.com/ideas/archive/2021/01/trump-is-banned-who-is-next/617622 (on file with the *Columbia Law Review*) (last updated Jan. 11, 2021) [hereinafter Douek, Trump Is Banned].

147. See id.

148. Jeremy Wright & Sajid Javid, Online Harms White Paper 42 (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf [https://perma.cc/BN3Q-R9RS].

149. Case C-131/12, Google Spain SL v. Agencia Española de Protección de Datos, ECLI:EU:C:2014:317 (May 13, 2014), http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&doclang=EN (on file with the *Columbia Law Review*).

cases.[150] Other legislation already enacted or proposed requires platforms to weigh social benefits of speech in deciding whether it falls within an exception from an obligation to remove certain types of content.[151] These include asking platforms to determine, for example, if content the legislation ordinarily would require platforms to take down should remain online because it is necessary for research or otherwise in the public interest.[152] Because platforms have global rules, laws passed in individual jurisdictions can create worldwide effects.[153]

These new or imminent laws, combined with changing societal expectations and norms, have forced platforms to abandon their professed neutrality and humility. They are now very much in the business of balancing rights and interests.

c. *Categorical Incoherence*. — The exposure of categories to these pressures and the course of time was harsh. The once simple categories governing platform speech have proliferated and broken down into ever-finer categories plagued by exceptions.[154] This follows the course of First Amendment jurisprudence, which "has become only more intricate, as categories have multiplied, distinctions grown increasingly fine, and exceptions flourished and become categories of their own."[155] Trying to

150. See, e.g., Daphne Keller, The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation, 33 Berkeley Tech. L.J. 287, 290 (2018) (noting that the court ordered Google to balance the right of a claimant seeking to delist with the right of other users to find information online); Robert C. Post, Privacy, Speech, and the Digital Imagination, *in* Free Speech in the Digital Age, supra note 122, at 104, 113 ("[T]he right to be forgotten can be successfully asserted only if the 'harm' to a data subject is balanced against the 'interest of the general public.'" (quoting *Google Spain*, ECLI:EU:C:2014:317, ¶¶ 81, 99)).

151. See, e.g., Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online, at 24, COM (2018) 0640 final (Sept. 12, 2018), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:52018PC0640 [https://perma.cc/8RM7-VZJ7] (stating that providers should consider the importance of freedom of expression in society); Douek, Nerd Harder, supra note 34, at 42 ("[W]riting laws to create social media reform is hard and involves difficult trade-offs."); Sandra Schmitz-Berndt & Christian M. Berndt, The German Act on Improving Law Enforcement on Social Networks: A Blunt Sword? 11–14 (Dec. 14, 2018) (unpublished working paper), https://ssrn.com/abstract=3306964 (on file with the *Columbia Law Review*) (noting that different countries' laws require platforms to regulate speech to varying extents).

152. Douek, Nerd Harder, supra note 34, at 46.

153. Danielle Keats Citron, Extremist Speech, Compelled Conformity, and Censorship Creep, 93 Notre Dame L. Rev. 1035, 1039–40 (2018).

154. For an indication of how these rule sets have expanded, see Community Standards, Facebook, https://www.facebook.com/communitystandards [https://perma.cc/SZE8-N4CU] (last visited Oct. 23, 2020); The Twitter Rules, Twitter, https://help.twitter.com/en/rules-and-policies/twitter-rules [https://perma.cc/R77G-NCRM] [hereinafter Twitter, Twitter Rules] (last visited Oct. 23, 2020); YouTube's Community Guidelines, YouTube, https://creatoracademy.youtube.com/page/course/community-guidelines [https://perma.cc/D483-2MAW] (last visited Oct. 23, 2020).

155. Elena Kagan, Private Speech, Public Purpose: The Role of Governmental Motive in First Amendment Doctrine, 63 U. Chi. L. Rev. 413, 515 (1996).

maintain categories has led to "typically tacit (and therefore baffling) distortions of the categories themselves."[156] In the words of Justice Stevens, "[E]fforts at categorization inevitably give rise only to fuzzy boundaries . . . . The quest for doctrinal certainty through the definition of categories and subcategories is . . . destined to fail."[157]

This experience was only accelerated in the context of content moderation due to the sheer volume of cases confronted across many more diverse cultures. Facebook's policy on female nipples, for example, now includes exceptions for breastfeeding and an expanding list of other specific contexts: "birth giving and after-birth moments, health-related situations (for example, post-mastectomy, breast cancer awareness or gender confirmation surgery) or an act of protest."[158] It is rule-like, but it is hard to describe this as a simple unitary category of adult nudity anymore.

ProPublica's exposé on Facebook's hate speech policy is another prominent example of categorical incoherence. In its attempt to devise a categorical policy that reduced discretion and value judgments on the part of frontline content moderators, Facebook's policy protected white men, but not Black children, from hate speech on the basis that gender was a protected category while childhood was not.[159] Rigid categories, unmoored from the justification for the rule, can appear baffling. Facebook's hate speech policy now has three separate "Tiers," and each tier has subcategories with very specific examples.[160] For instance, "cursing" is listed and defined in "Tier 2,"[161] with a specificity that makes quoting it likely inappropriate for the pages of the *Columbia Law Review*.

Reddit is another platform that exemplifies this general trend. It long resisted the idea of proactive moderation and refused to remove all but the worst content on its platform, earning a reputation as a "cesspool of racism."[162] But cultural pressure caused Reddit's CEO to struggle "with balancing [his] values as an American, and around free speech and free expression, with [his] values and the company's values around common

156. Greene, supra note 51, at 33.

157. R.A.V. v. City of St. Paul, 505 U.S. 377, 426 (1992) (Stevens, J., concurring in the judgment).

158. Community Standards: Adult Nudity and Sexual Activity, Facebook, https://www.facebook.com/communitystandards/adult_nudity_sexual_activity [https://perma.cc/H6MR-T89K] (last visited Oct. 23, 2020).

159. Julia Angwin & Hannes Grassegger, Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children, ProPublica (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/T5XW-TC5Q].

160. Community Standards: Hate Speech, Facebook, https://www.facebook.com/communitystandards/hate_speech [https://perma.cc/EM7U-7G6Y] (last visited Oct. 23, 2020).

161. Id.

162. Kevin Roose, Reddit's C.E.O. on Why He Banned 'The_Donald' Subreddit, N.Y. Times (June 30, 2020), https://www.nytimes.com/2020/06/30/us/politics/reddit-bans-steve-huffman.html (on file with the *Columbia Law Review*).

human decency."[163] The company therefore rolled out a new hate speech policy.[164] An initial attempt at providing a more comprehensive definition of categories for what it would consider hate speech was incoherent and underspecified (including an exemption for groups "who are in the majority") and required immediate revision.[165] The amended rule is broader and requires an "all things considered" contextual evaluation.[166]

So, the question is: "[W]hat comes next in a free speech world without firm categories[?]"[167] With the destabilization of the posts-as-trumps categorical framework came the rise of the globally dominant, alternative lens for rights adjudication: proportionality.

d. *Global Platforms and Global Constitutionalism.* — Proportionality is the globally dominant form of rights adjudication, with the notable exception of the United States.[168] It should therefore be unsurprising that this frame came to implicitly dominate the framing of content moderation issues as the major platforms became global services. Under a categorical approach, the inquiry focuses on the identification of a speech right and its boundaries. A proportionality approach, by contrast, accepts that speech rights can be limited for legitimate purposes, provided any infringement is proportionate. Rather than playing taxonomist, the decisionmaker must identify and weigh multiple interests like a grocer or epidemiologist.

---

163. Newton, Reddit Bans, supra note 3 (internal quotation marks omitted) (quoting Steve Huffman).

164. Promoting Hate Based on Identity or Vulnerability, Reddit Help, http://reddit.zendesk.com/hc/en-us/articles/360045715951 [https://perma.cc/JAF6-3P8K] (last visited Nov. 16, 2020).

165. Adriana Stephan, Comparing Platform Hate Speech Policies: Reddit's Inevitable Evolution, Stan. Freeman Spogli Inst. for Int'l Stud. (July 8, 2020), https://fsi.stanford.edu/news/reddit-hate-speech [https://perma.cc/ZMT3-NB62].

166. Promoting Hate Based on Identity or Vulnerability, supra note 164.

167. Ashutosh Bhagwat, Free Speech Categories in the Digital Age, *in* Free Speech in the Digital Age, supra note 122, at 88–89.

168. See, e.g., Barak, Constitutional Rights and Their Limitations, supra note 73, at 206; Vicki C. Jackson & Mark V. Tushnet, Introduction to Proportionality: New Frontiers, New Challenges 1, 1 (Vicki C. Jackson & Mark V. Tushnet eds., 2017); Stone Sweet & Mathews, Proportionality Balancing and Constitutional Governance, supra note 14, at 96; Evelyn Douek, All Out of Proportion: The Ongoing Disagreement About Structured Proportionality in Australia, 47 Fed. L. Rev. 551, 552 (2019) [hereinafter Douek, All Out of Proportion] (noting that though the proportionality approach is globally dominant, its place in Australian law has been "contentious"); Greene, supra note 51, at 58–59; Jackson, supra note 15, at 3096; Bernard Schlink, Proportionality in Constitutional Law: Why Everywhere but Here?, 22 Duke J. Compar. & Int'l L. 291, 297–98 (2012); Alec Stone Sweet & Jud Mathews, Proportionality Balancing and Global Constitutionalism, 47 Colum. J. Transnat'l L. 72, 74 (2008); Grégoire C.N. Webber, Proportionality, Balancing, and the Cult of Constitutional Rights Scholarship, 23 Can. J.L. & Juris. 179, 180 (2010) [hereinafter Webber, The Cult of Constitutional Rights Scholarship].

As discussed further below,[169] proportionality is not *mere* balancing, and it is the failure to fully realize this that has led to widespread disillusionment with the current approach to online speech governance. Online speech governance has adopted the globally dominant form of rights adjudication without paying heed to the many lessons that other jurisdictions' experiences with proportionality can teach. It was likely inevitable that proportionality would infuse online speech governance as ever-increasing majorities of online speakers and platform decisionmakers came from jurisdictions where this approach is the norm and the American categorical approach is seen as the outlier.

2. *The Pros of Proportionality.* — So far, this Article has explored *why* online speech governance increasingly coalesced around a proportionality approach. This section argues that this development is not merely understandable, but also salutary. As the internet ages, proportionality is a more mature approach to resolving the many conflicts created by the collision of varying interests online. There are three main benefits: (a) it explicitly acknowledges interests other than the individual speech right, and thereby dignifies those interests and the importance of evaluating them in their particular context; (b) it is transparent about the value judgments inherent in constructing a system of freedom of expression; and (c) it encourages and rationalizes remedial flexibility, rather than a binary "take-down/leave-up" paradigm of content moderation.

a. *Acknowledging Multiple Interests.* — A proportionality framework openly acknowledges competing rights and interests, and provides tools for evaluating them in the context of particular controversies. The core strength of proportionality is that it allows conflicts between competing values to be resolved by reference to the specific circumstances, without creating abstract and definitive value hierarchies.[170] Indeed, for critics, this contextual evaluation rather than adherence to clear rules is a core *weakness* of proportionality as a method because it undermines predictability and uniformity.[171]

Recognizing multiple competing interests and their context of course changes substantive rules, but it also has relational effects for stakeholders in conflicts: Acknowledging competing interests dignifies them. As Professor Jamal Greene argues, "Because the rights-as-trumps frame cannot accommodate conflicts of rights, it forces us to deny that our opponents have them. When rights are trumps, they favor rhetoric over judgment, simplicity over context, homogeneity over diversity."[172] Those

169. See infra section III.B.

170. Niels Petersen, Proportionality and Judicial Activism: Fundamental Rights Adjudication in Canada, Germany and South Africa 38 (2017).

171. See, e.g., Webber, The Cult of Constitutional Rights Scholarship, supra note 168, at 199–200. Section III.B discusses this idea further.

172. Greene, supra note 51, at 34.

competing interests might not prevail, but recognizing them can itself increase perception of the legitimacy of the process.[173]

This is especially important for global platforms. One of the central criticisms in recent years has been major platforms' failure to appreciate different demands of varying contexts.[174] Proportionality provides an analytical framework through which decisionmakers can take notice of and evaluate local interests. A categorical approach says that "this kind of speech is (im)permissible"; a proportionality approach can say that "*in this context*, this kind of speech does greater/less harm than in other contexts and so should be treated differently." This contextual nature of proportionality adjudication is why the adoption of international human rights norms (a proportionality-based system) as the basis for platform rules would not result in universal, homogenous platform rules across every market, but would instead allow for more sensitive adaptation to local context.[175]

b. *Transparency.* — In proportionality's ideal form, explicit acknowledgment of the interests being balanced also furthers the rule of law values by making the actual basis of decisions transparent.[176] Value judgments are unavoidable, even for categorical decisionmakers.[177] Proportionality acknowledges this truth and makes such judgments explicitly, rather than denying their existence and smuggling them in through categorization that only becomes ever more intricate and distorted.[178] It is the very

---

173. See Facebook Data Transparency Advisory Group, supra note 40, at 33–34 ("[J]udgments about legitimacy do not depend primarily on whether authorities give them favorable outcomes . . . . Rather, judgments about legitimacy are more strongly swayed by the processes and procedures by which authorities use their authority . . . ."); see also Tyler, supra note 40, at 286 (discussing how procedural justice enhances legitimacy).

174. See, e.g., Kaye, Speech Police, supra note 59, at 117 ("The companies are not built to moderate content at global scale. They often alienate and flatten the cultures across the markets where they operate."); Vaidhyanathan, supra note 118, at 27 ("Facebook has universalizing tendencies and embodies a globalist ambition. But it does not work the same way in Phnom Penh as it does in Philadelphia."); Chinmayi Arun, Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms (Jan. 23, 2018) (unpublished manuscript), https://ssrn.com/abstract=3108238 (on file with the *Columbia Law Review*) ("Harmful speech can be hyper-localised . . . such that even people from the same state might not understand. This is the sort of thing that . . . a global corporation might miss completely.").

175. See Jen Patha Howell, The Lawfare Podcast: David Kaye on Policing Speech Online, Lawfare (Dec. 5, 2019), https://www.lawfareblog.com/lawfare-podcast-david-kaye-policing-speech-online (on file with the *Columbia Law Review*).

176. See Facebook Data Transparency Advisory Group, supra note 40, at 38–39.

177. See supra section I.B.1.a.

178. See Barak, Constitutional Rights and Their Limitations, supra note 73, at 488 ("[P]roportionality makes transparent legal decisions. Categorization tends to be less transparent. The reasons underlying the categorical choice are typically not made explicit."); Petersen, supra note 170, at 58 ("[C]ategorical forms of argumentation either neglect the demands of fit or are less transparent than arguments based on balancing."); Douek, All Out of Proportion, supra note 168, at 557 ("By separating out the elements, structured proportionality seeks to make unavoidable value judgments more explicit.").

openness of interest balancing that makes it a poor vehicle for importing illegitimate considerations.[179] Of course, this openness is the aspiration of proportionality, but "[a]ny test can be applied badly or well."[180] As discussed below,[181] so far this test has been applied very badly indeed in online speech governance. The important point is that proportionality is a concession to realism: Decisionmakers *are* balancing competing interests when they construct rules; proportionality is candid about it. Such candor, aside from being intrinsically valuable, can operate as a constraint by making the judgments open to scrutiny.

c. *Remedial Flexibility.* — Proportionality also facilitates a more nuanced approach to remedies. By contrast to a posts-as-trumps framework—which lends itself to binary remedies (Against the rules? Take it down! Otherwise? Leave it up!)—a proportionality framework goes hand-in-hand with remedial flexibility.[182] Indeed, proportionality requires that any restriction on freedom of expression should only be that which is "necessary," in the sense that it is the least restrictive means of achieving the desired outcome.[183] This entails looking at the full range of options between leaving content up or taking it down.

The availability of these options is an underappreciated aspect of the online environment: Platforms can develop far more nuanced remedies than have traditionally been available to governments.[184]

Indeed, platforms are beginning to take advantage of this wider range of remedial options, further evidencing the implicit rise of proportionality as a governing tenet. Zuckerberg has described how Facebook is increasingly addressing "borderline content" (content approaching the line for what is disallowed, but not crossing it) by "reducing its distribution and virality" rather than removing it.[185] Fact-checked content is flagged with

179. See Petersen, supra note 170, at 189.

180. Jeremy Kirk, Constitutional Guarantees, Characterisation and the Concept of Proportionality, 21 Melb. U. L. Rev. 1, 63 (1997).

181. See infra section II.A.

182. See Greene, supra note 51, at 115–19; Jackson, supra note 15, at 3178 (advocating for an approach to remedies in constitutional litigation that focuses not "only on the nature of the classification" but also "on the relative nature of the harm" and its relationship to the governmental interests at stake, which "would allow courts to hold legislatures accountable without invalidating most legislation").

183. Barak, Constitutional Rights and Their Limitations, supra note 73, at 317–39.

184. See David Kaye, Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, at 16–17, U.N. Doc. A/74/486 (Oct. 9, 2019) [hereinafter Kaye, Freedom of Opinion and Expression].

185. Mark Zuckerberg, A Blueprint for Content Governance and Enforcement, Facebook (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 [https://perma.cc/MDD9-8CGU] [hereinafter Zuckerberg, Blueprint for Content Governance].

context, rather than censored.[186] Twitter adds notice screens and reduces circulation of public figures' tweets that breach the site's rules rather than taking them down entirely.[187] Facebook has announced it will do the same for certain posts in the public interest.[188] These intermediate solutions leverage the affordances of platform architecture and move away from the take-down/leave-up binary in order to "strike the right balance between enabling free expression, fostering accountability, and reducing the potential harm caused by [content]."[189] That is, these remedies are a more *proportionate* response than removing content entirely.

There are many more such possible interventions, including adding transparency about the identity of users or pages,[190] labeling manipulated media,[191] demonetizing content,[192] restricting the ability to share content or otherwise adding friction,[193] geoblocking content to particular regions,[194] enabling counter-messaging,[195] or nudging users toward authoritative information.[196] The possibilities are limited more by imagination than by any immutable characteristics of platforms.

---

186. How Is Facebook Addressing False Information Through Independent Fact-Checkers?, Facebook Help Ctr., https://www.facebook.com/help/1952307158131536?helpref=faq_content [https://perma.cc/ZV89-B7A9] (last visited Oct. 23, 2020).

187. Defining Public Interest on Twitter, Twitter Safety (June 27, 2019), https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html [https://perma.cc/BWA9-EWZ5].

188. Mark Zuckerberg, Facebook (June 26, 2020), https://www.facebook.com/zuck/posts/10112048980882521 [https://perma.cc/QPV6-PEQK].

189. Defining Public Interest on Twitter, supra note 187.

190. Anita Joseph & Georgina Sheedy-Collier, Making Pages and Accounts More Transparent, Facebook: Newsroom (Apr. 22, 2020), https://about.fb.com/news/2020/04/page-and-account-transparency [https://perma.cc/XZ59-NW5W].

191. Synthetic and Manipulated Media Policy, Twitter Help Ctr., https://help.twitter.com/en/rules-and-policies/manipulated-media [https://perma.cc/A5K4-UZ5V] (last visited Oct. 22, 2020).

192. Robyn Caplan & Tarleton Gillespie, Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy, Social Media + Society, Apr.–June 2020, at 2; Advertiser-Friendly Content Guidelines, YouTube Help, https://support.google.com/youtube/answer/6162278?hl=en [https://perma.cc/2N6E-J2PJ] (last visited Oct. 22, 2020).

193. Keeping WhatsApp Personal and Private, WhatsApp Blog (Apr. 7, 2020), https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private [https://perma.cc/JGJ7-YG82].

194. Dan Jerker B. Svantesson, Solving the Internet Jurisdiction Puzzle 203 (2017).

195. Daniel Kreiss & Matt Perault, Opinion, Four Ways to Fix Social Media's Political Ads Problem—Without Banning Them, N.Y. Times (Nov. 16, 2019), https://www.nytimes.com/2019/11/16/opinion/twitter-facebook-political-ads.html (on file with the *Columbia Law Review*).

196. The Four Rs of Responsibility, Part 2: Raising Authoritative Content and Reducing Borderline Content and Harmful Misinformation, YouTube Off. Blog (Dec. 3, 2019), https://youtube.googleblog.com/2019/12/the-four-rs-of-responsibility-raise-and-reduce.html [https://perma.cc/9GFD-RY7Y]; Rosen, Facebook COVID-19 Update, supra note 4; WHO to Livestream on TikTok, TikTok: Newsroom (Mar. 16, 2020), https://newsroom.tiktok.com/en-us/who-to-livestream-on-tiktok [https://perma.cc/7ES8-YMC6].

A categorical frame makes these responses seem incoherent: Once content falls within a "problematic" category, half-measures seem like simple lack of conviction.[197] There is no frame by which to justify treating content within the same category differently. But a proportionality framework legitimates and rationalizes these tools by acknowledging the competing interests at stake.

Another important requirement of proportionality is the need for different procedural protections for different types of content, consequences, and remedies.[198] It may be acceptable, for example, to afford lesser procedural protections when attaching a link to more information for every post that mentions 5G than removing an entire account. Because kicking a user off a platform entirely can have significant consequences for that individual,[199] robust procedural protections should be afforded.

* * *

I have painted an idealistic view of the proportionality paradigm so far, but there are very real deficits in the way this approach has been applied in practice.[200] Before turning to examine those deficits, however, I explore the other key dimension along which early platform governance came under strain and has had to adapt: moving from an individualistic approach to a probabilistic one.

C. *The Unavoidability of Probability*

Speech rights are traditionally individualistic: Free speech history and jurisprudence typically focus on stories of particular individuals and even particular utterances.[201] Constitutional cases tell iconic stories with identifiable protagonists who said, or wore, now-iconic words or phrases.[202]

197. See, e.g., Zuckerberg, Blueprint for Content Governance, supra note 185 ("One common reaction is that rather than reducing distribution, we should simply move the line defining what is acceptable.").

198. See Evelyn Douek, Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation 10 (Hoover Inst. Aegis Paper Series, Paper No. 1903, 2019) (noting that content regulation "needs to be determined contextually").

199. Kashmir Hill, Many Are Abandoning Facebook. These People Have the Opposite Problem., N.Y. Times (Aug. 22, 2019), https://www.nytimes.com/2019/08/22/business/reactivate-facebook-account.html (on file with the *Columbia Law Review*) (last updated Nov. 22, 2019).

200. See infra section III.B.1.

201. Owen M. Fiss, Free Speech and Social Structure, 71 Iowa L. Rev. 1405, 1408 (1986) ("For the most part, the Free Speech Tradition can be understood as a protection of the street corner speaker. An individual mounts a soapbox on a corner in some large city, starts to criticize governmental policy, and then is arrested for breach of the peace.").

202. Laura Weinrib, Rethinking the Myth of the Modern First Amendment, *in* The Free Speech Century 48, 48 (Lee C. Bollinger & Geoffrey R. Stone eds., 2019) (describing the importance of seminal First Amendment cases and their defendants in creating the modern myth of the First Amendment).

Positive human rights law similarly vests rights in individuals: Article 19 of the International Covenant on Civil and Political Rights begins, "*Everyone* shall have the right to freedom of expression."[203] But U.S. First Amendment doctrine has been especially resistant to silencing individuals in aid of some conception of the common good.[204] To be sure, there is a strong lineage of collectivist, structural, or instrumental theories of the First Amendment that focus on the value of speech for some greater purpose (most commonly self-government).[205] Actual doctrine, however, has been "largely hostile" to such agendas,[206] and collectivist theories of the First Amendment have not gained significant traction outside the academy. Alexander Meiklejohn's famous aphorism that "[w]hat is essential is not that everyone shall speak, but that everything worth saying shall be said"[207] has not won out in the marketplace of ideas. The exceptional doctrine of overbreadth epitomizes this approach, holding that laws will be invalidated even when they apply to a substantial amount of prohibitable speech if they would also overly infringe on protected speech.[208] In short, the idea that some speakers should tolerate having their protected speech silenced for some greater good is anathema to the First Amendment.

Each content moderation decision "look[s] like" a traditional speech case, and so lends itself to analogies with individual rights decisions.[209] This is why ideas like a content moderation "Supreme Court" (the original moniker of what is now known as the Facebook Oversight Board, an independent body set up to review Facebook's decisions) have intuitive appeal.[210] But the next section shows that adopting a highly individualistic lens is impractical for online speech governance: The sheer scale and

---

203. International Covenant on Civil and Political Rights art. 19, Dec. 16, 1966, 999 U.N.T.S. 171, 178 [hereinafter ICCPR] (emphasis added).

204. Rosenfeld, Hate Speech in Constitutional Jurisprudence, supra note 139, at 1541 ("[F]ree speech in the United States is shaped above all by individualism and libertarianism . . . .").

205. See Alexander Meiklejohn, Free Speech and Its Relation to Self-Government 3 (1948).

206. See Robert Post, Meiklejohn's Mistake: Individual Autonomy and the Reform of Public Discourse, 64 U. Colo. L. Rev. 1109, 1109 (1993); Morgan N. Weiland, Expanding the Periphery and Threatening the Core: The Ascendant Libertarian Speech Tradition, 69 Stan. L. Rev. 1389, 1404–12 (2017) ("The republican tradition arguably is less prominent in First Amendment law as compared to the liberal tradition . . . .").

207. Meiklejohn, supra note 205, at 25.

208. Richard H. Fallon Jr., Making Sense of Overbreadth, 100 Yale L.J. 853, 863 (1991).

209. See Margot E. Kaminski, Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability, 92 S. Cal. L. Rev. 1529, 1571 (2019) (acknowledging that questions about systemic accountability in algorithmic decisionmaking have been obscured by the fact that many decisions "look like" decisions that traditionally invoked questions of individual due process).

210. See Evelyn Douek, Facebook's New 'Supreme Court' Could Revolutionize Online Speech, Lawfare (Nov. 19, 2018), https://www.lawfareblog.com/facebooks-new-supreme-court-could-revolutionize-online-speech [https://perma.cc/S2DH-UBZ6].

diversity of online speech belies thinking through the traditional frame of each individual case. Content moderation is inherently systemic. Content moderation systems do not promise to get every individual speech decision right; they are designed to increase the probability that most decisions will be right most of the time and when the system errs, it does so in a preferred direction. A systemic approach accepts the inevitability of errors and factors them into governance design.

This section starts by illustrating the impossibility of an individualistic approach to online speech governance. Next, it turns to the way that platforms have dealt with this problem. Artificial intelligence (AI) tools have become indispensable for dealing with the unfathomable firehose of online speech. Such tools are fundamentally probability-based in a way that seems at odds with an individualistic understanding of speech rights. The final section argues that probabilistic analysis is therefore an unavoidable and pragmatic principle of the future of online speech governance.

1. *Content Moderation Is Impossible.* — Scale is major platforms' Prime Directive: both what they strive to achieve, and (having done so) what everything must be built around. For speech governance, the change wrought by the platforms' scale is a shift in kind, not degree.

The scale is truly unfathomable. In the last quarter of 2020, Facebook "took action" on over 105 million pieces of content (an average of about 1.1 million actions *per day*) and Instagram on over 35 million.[211] YouTube, which has 500 hours of new video uploaded every *minute*,[212] removed over 9.3 million videos (over 100,000 per day).[213] In the first half of 2020, Twitter dealt with reports against over 12.4 million unique accounts for potential violations of the Twitter rules and took action against over 1.9 million.[214] As the former Chief Security Officer of Facebook Alex Stamos told Congress, these numbers only represent cases where platforms took action (or in Twitter's case, had reports): "[T]he overall number of decisions considered, including those where no action was taken, is much higher."[215]

---

211. Guy Rosen, Community Standards Enforcement Report, February 11 Edition, Facebook (Feb. 11, 2021), https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020 [https://perma.cc/V2CW-4ZGP]. This figure excludes removals of fake accounts and spam, to confine the number to substantive rule determinations. If these categories were included, the figure would be over 2.4 *billion* actions in the quarter for Facebook.

212. The Tricky Task of Policing YouTube, Economist (May 4, 2019), https://www.economist.com/briefing/2019/05/04/the-tricky-task-of-policing-youtube (on file with the *Columbia Law Review*).

213. YouTube Community Guidelines Enforcement, Google Transparency Rep., https://transparencyreport.google.com/youtube-policy/removals?hl=en [https://perma.cc/Q5BP-ZS2F] (last visited Jan. 26, 2021).

214. Rules Enforcement, Twitter Transparency, https://transparency.twitter.com/en/twitter-rules-enforcement.html https://perma.cc/6F85-MMJY] (last visited Jan. 26, 2021).

215. Artificial Intelligence and Counterterrorism: Possibilities and Limitations: Hearing Before the Subcomm. on Intel. & Counterterrorism of the H. Comm. on Homeland Sec.,

It is not just *hard* to get content moderation right at this scale; it is *impossible*. Speech decisions in any context are difficult—courts get them wrong all the time. Even if there were clearly "right" answers, and even if platform content moderators had unlimited time and resources to devote to every decision, the inevitability of error means that the sheer number of decisions would still result in a very large number of mistakes.[216] Indeed, "given the sheer enormity of the undertaking, most platforms' definition of success includes failing users on a regular basis."[217] Put another way, "To Err Is Platform."[218] Accordingly, "moderation at the major platforms is *as much a problem of logistics as a problem of values*."[219]

2. *Tools and Systems*. — At such unfathomable scale, the best a system of online speech governance can hope to do is minimize error. In platforms' early days, the primary mechanism for doing this was hiring more content moderators—armies of contractors reviewing content against internal guidelines that break "complex philosophical ideals about what constitutes harassment, hate, or truth into small components that are more likely to be interpretable."[220] But this has always had limits,[221] and until recently there was "the assumption that platforms could not, realistically, monitor user speech on an ongoing basis."[222]

But the capacity to monitor and enforce online speech has changed dramatically in the past few years, primarily as a result of increased

---

116th Cong. 9 (2019) (statement of Alexander Stamos, Adjunct Professor and Program Director, Stanford Internet Observatory, and Former Chief Security Officer, Facebook).

216. Tech journalist Mike Masnick has coined this idea "Masnick's Impossibility Theorem." See Mike Masnick, Masnick's Impossibility Theorem: Content Moderation at Scale Is Impossible to Do Well, Techdirt (Nov. 20, 2019), https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml [https://perma.cc/52X2-BGEE] ("[W]hen you're doing content moderation at scale, the scale part means that even if you're very, very, very, very good, you will still make a ridiculous number of mistakes in absolute numbers every single day.").

217. Gillespie, Custodians of the Internet, supra note 59, at 197.

218. James Grimmelman, To Err Is Platform, Knight First Amend. Inst. (Apr. 6, 2018), https://knightcolumbia.org/content/err-platform [https://perma.cc/5S27-UFVE].

219. Gillespie, Custodians of the Internet, supra note 59, at 116 (emphasis added).

220. Robyn Caplan, Content or Context Moderation? 23–24 (2018), https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf [https://perma.cc/7K53-PTC7].

221. See Tarleton Gillespie, Content Moderation, AI, and the Question of Scale, Big Data & Soc'y, July–Dec. 2020, at 1, 1 [hereinafter Gillespie, Content Moderation] (quoting Jack Dorsey, Chief Executive Officer of Twitter, as saying "there are no amount of people that can actually scale this," only algorithms can (internal quotation marks omitted)).

222. Daphne Keller & Paddy Leerssen, Facts and Where to Find Them: Empirical Research on Internet Platforms and Content Moderation, *in* Social Media and Democracy: The State of the Field and Prospects for Reform 220, 224 (Nathaniel Persily & Joshua A. Tucker eds., 2020).

"algorithmic moderation"[223]—the use of automated techniques to classify content and apply a content moderation outcome to it. Such algorithmic moderation has grown radically in the last few years alone,[224] and is "increasingly taking the place of human content moderation."[225] Indeed, the rise of these tools has been so rapid that Professor Kate Klonick's seminal 2018 article on content moderation only briefly discusses automated moderation.[226] By contrast, in Mark Zuckerberg's first appearance before Congress in April 2018, he referred to AI more than thirty times, saying that over the next decade it would become "the scalable way to identify and root out most of this harmful content."[227] Algorithmic moderation is thus a stark manifestation of Professor Lawrence Lessig's prediction over a decade ago that code would become an ever more important regulator.[228] Yet, to date, "the role of automation in this context has received scant scholarly attention."[229]

Algorithmic moderation allows unprecedented amounts of speech to be subject to enforcement action at unprecedented speeds.[230] Platforms tout ever greater percentages of removed content caught "proactively" by AI tools,[231] even in those categories of content, like hate speech, that

223. Robert Gorwa, Reuben Binns & Christian Katzenbach, Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance, Big Data & Soc'y, Jan.–June 2020, at 1, 3.

224. See, e.g., Zuckerberg, Blueprint for Content Governance, supra note 185 ("Moving from reactive to proactive handling of content at scale has only started to become possible recently because of advances in artificial intelligence . . . .").

225. York & Zuckerman, supra note 55, at 149; see also Daphne Keller, Observations on Speech, Danger, and Money 5 (2018), https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf [https://perma.cc/4KP2-8AME] [hereinafter Keller, Observations] ("[T]o handle the expanded volume of takedowns, both major notifiers and major platforms rely increasingly on automation rather than human review.").

226. See Klonick, supra note 49, at 1635–37, 1635 n.261.

227. Drew Harwell, AI Will Solve Facebook's Most Vexing Problems, Mark Zuckerberg Says. Just Don't Ask When or How., Wash. Post (Apr. 11, 2018), https://www.washingtonpost.com/news/the-switch/wp/2018/04/11/ai-will-solve-facebooks-most-vexing-problems-mark-zuckerberg-says-just-dont-ask-when-or-how (on file with the *Columbia Law Review*); see also Evelyn Douek, Zuckerberg's New Hate Speech Plan: Out with the Court and In with the Code, Lawfare (Apr. 14, 2018), https://www.lawfareblog.com/zuckerbergs-new-hate-speech-plan-out-court-and-code [https://perma.cc/TS48-GW27] ("The plan epitomizes technological optimism: in five to ten years, Zuckerberg said, he expects artificial intelligence will able to proactively monitor posts for hateful content. In the meantime? Facebook is hiring more human content moderators.").

228. Lawrence Lessig, Code: Version 2.0, at 125–37 (2006).

229. Bloch-Wehba, Automation in Moderation, supra note 45 (manuscript at 4).

230. Daphne Keller, Who Do You Sue? State and Platform Hybrid Power over Online Speech 1 (2019), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf [https://perma.cc/8Q2E-Q7S7].

231. See supra notes 191–194 and accompanying text; see also Emine Saner, YouTube's Susan Wojcicki: 'Where's the Line of Free Speech—Are You Removing Voices that Should Be Heard?', Guardian, (Aug. 10, 2019), https://www.theguardian.com/technology/2019/aug/10/youtube-susan-wojcicki-ceo-where-line-removing-voices-heard [https://perma.cc/

require highly contextual evaluation and are therefore hard to moderate at all, let alone with AI.[232] That platforms now *can* police so much content has led to expectations that they *ought* to: There is a feedback loop in which the use of algorithmic content moderation both responds to and causes government and societal demands that platforms take greater control of the content on their services.

Wu suggests that these tools make decisions "in the shadow of the law, but . . . [are] not compelled by it, and the decisions made by the software are now more important than the law."[233] Increasingly, though, these tools *are* being compelled by law—if not explicitly, then implicitly by virtue of requirements that would be impossible to comply with without AI.[234] Proactive monitoring and filtering obligations are increasingly being imposed on intermediaries across more types of content by laws around the world.[235] For example, a recent Court of Justice of the European Union order requires platforms to use "automated search tools and technologies" to remove content deemed identical or equivalent to content declared to be illegal.[236] The new EU copyright directive requires platforms to make best efforts to block uploads of copyrighted material.[237] Australia passed a law requiring platforms to remove "abhorrent violent

---

L3AE-F88X] ("The machines get more and more sophisticated as we get better at identifying that content, which means we can find it faster than before. Over 75% of the content that we're removing with machines, we find and remove without even a single view." (internal quotation marks omitted) (quoting Susan Wojcicki, Chief Executive Officer of YouTube)). I use the term "AI tools" generically and broadly to mean any automated software used for decisionmaking.

232. See, e.g., Will Knight, Facebook's Head of AI Says the Field Will Soon 'Hit the Wall', WIRED (Dec. 4, 2019), https://www.wired.com/story/facebooks-ai-says-field-hit-wall [https://perma.cc/QZ6Q-Y7Y4] ("[T]here's been a lot of progress in the field of language, allowing us a much more refined understanding of interactions . . . . We can understand . . . if it's hate speech, or if it's just a joke. By no measure is it a solved problem, but there's clear progress being made.").

233. Wu, Will Artificial Intelligence Eat the Law, supra note 22, at 2007. Wu is referring to copyright infringement detection technology, but the observation applies more broadly.

234. See Gorwa et al., supra note 223, at 2 ("Under recent regulatory measures . . . platforms are increasingly being bound to a very short time window for content takedowns that effectively necessitates their use of automated systems to detect illegal or otherwise problematic material proactively and at scale.").

235. See Giancarlo Frosio & Sunimal Mendis, Monitoring and Filtering: European Reform or Global Trend?, *in* Oxford Handbook of Online Intermediary Liability 544, 547 (Giancarlo Frosio ed., 2020) [hereinafter Online Intermediary Liability].

236. See Case C-18/18, Glawischnig-Piesczek v. Facebook Ireland, ECLI:EU:C:2019:821, ¶¶ 46, 55 (Oct. 3, 2019), http://curia.europa.eu/juris/document/document.jsf?text=&docid=218621&doclang=EN (on file with the *Columbia Law Review*).

237. See Julia Reda, Berkman Klein Ctr., Why Americans Should Worry About the New EU Copyright Rules, Medium (Dec. 20, 2019), https://medium.com/berkman-klein-center/why-americans-should-worry-about-the-new-eu-copyright-rules-97800be3f8fc [https://perma.cc/JWB3-ZZXR].

material . . . expeditiously."[238] The timeframes implicit in this demand make anything but automated moderation likely to be inadequate to meet it.[239] Other laws similarly stack the deck in incentivizing platforms to automate their operations.[240] AI tools are increasingly necessary to meet public and regulatory expectations of content moderation.[241]

This is the fairytale version of the story: Faced with unprecedented and ungovernable volumes of speech and societal and regulatory demands that they moderate more responsibly, platforms developed magical AI tools that tamed the ocean of online content. But this is not the reality. The inevitability of error has not disappeared; it has merely changed.

Platforms' claims about the accuracy of algorithmic moderation are unverified.[242] Even taking them at face value, it is obvious that AI tools do not remove the inevitability of error. AI tools may make different errors than humans and may even err more predictably, but they still make errors.[243] Too often, the label "AI" is used to mask the bluntness and vulnerabilities of these tools.

There are two types of automated tools used in commercial content moderation: matching systems, which compare new posts against a database of preclassified content, and predictive systems, which aim to classify new posts as against platform rules.[244] Both make errors, but different kinds.[245] Matching systems involve both false positives (like when a match is recorded despite context changing the content's meaning, such as

238. Douek, Nerd Harder, supra note 34, at 49–53 (internal quotation marks omitted) (quoting Christian Porter, Austl. Att'y Gen., Address at the Reading of the Criminal Code Amendment Bill 1850 (Apr. 4, 2019), https://parlinfo.aph.gov.au/parlInfo/genpdf/chamber/hansardr/84457b57-5639-432a-b4df-68b704cb3563/0032/hansard_frag.pdf;fileType=application%2Fpdf (on file with the *Columbia Law Review*)).

239. See id. at 45 (noting that timelines for removal will likely be measured in "hours and minutes, rather than days").

240. See, e.g., Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online, supra note 151, at 13 (proposing obligations on service providers to remove terrorist content within one hour of receiving a removal order); Heidi Tworek & Paddy Leerssen, An Analysis of Germany's NetzDG Law 2 (Apr. 15, 2019) (unpublished working paper), https://www.ivir.nl/publicaties/download/NetzDG_Tworek_Leerssen_April_2019.pdf [https://perma.cc/Q6RH-8DAA] ("If the content [in Germany] is 'manifestly unlawful,' platforms must remove it within 24 hours. Other illegal content must be taken down within 7 days.").

241. Gorwa et al., supra note 223, at 2.

242. See Gillespie, Content Moderation, supra note 221, at 3 ("[R]ecent claims by platforms of successful automated moderation are overstated.").

243. Richard M. Re & Alicia Solow-Niederman, Developing Artificially Intelligent Justice, 22 Stan. Tech. L. Rev. 242, 255–56 (2019) ("Of course, AI adjudication will still make 'mistakes,' however defined."); see also Zuckerberg, Blueprint for Content Governance, supra note 185 ("[W]hile computers are consistent at highly repetitive tasks, people are not always as consistent in their judgements.").

244. Gorwa et al., supra note 223, at 3.

245. For a helpful, practical overview of filtering error types, see generally Daphne Keller, Dolphins in the Net: Internet Content Filters and the Advocate General's *Glawischnig-Piesczek v. Facebook Ireland* Opinion (2019).

terrorist footage being used in news reporting) and false negatives (like when a match is missed because of some immaterial alteration to the content, such as the at least 800 visually distinct variants of the Christchurch Massacre uploaded to Facebook altered to evade detection[246]). Predictive systems have more generalized and potentially problematic errors. While there is little to no information about the tools platforms actually use, the best evidence is that they are brittle and easily fooled.[247] Research about the biases and blind spots of predictive language tools is widespread and mounting.[248] But such biases and blind spots might be better described as errors that occur when AI tools go *wrong*. Even when AI tools get it *right*, state of the art tools still make errors.[249]

The unavoidable truth is these tools are deployed with the full expectation that they will get decisions wrong. Given the volume of decisions involved, this cannot be fully mitigated by simply hiring more content moderators, even as that remains an essential demand for platforms to meet.

This Article has focused on moderation tools that detect and remove content, but the probability framework applies more broadly too. Facebook, for example, claims success when it reduces the impressions received by fact-checked stories by eighty percent.[250] WhatsApp touts a seventy-percent reduction in forwarding of viral messages as evidencing real progress in the fight against misinformation.[251] YouTube declared

---

246. See Chris Sonderby, Update on New Zealand, Facebook: Newsroom (Mar. 18, 2019), https://newsroom.fb.com/news/2019/03/update-on-new-zealand [https://perma.cc/TH5K-AJG5].

247. See Hossein Hosseini, Sreeram Kannan, Baosen Zhang & Radha Poovendran, Deceiving Google's Perspective API Built for Detecting Toxic Comments 1 (Feb. 27, 2017) (unpublished manuscript), https://arxiv.org/pdf/1702.08138.pdf [https://perma.cc/USQ5-SX69] (demonstrating how Google's machine learning project used to detect toxic language can be easily tricked).

248. See, e.g., Natasha Duarte, Emma Llanso & Anna Loup, Ctr. for Democracy & Tech., Mixed Messages? The Limits of Automated Social Media Content Analysis 8–9 (2017), https://cdt.org/wp-content/uploads/2017/11/Mixed-Messages-Paper.pdf [https://perma.cc/H45H-YRD6]; Facebook's Civil Rights Audit, supra note 25, at 76–82; Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi & Noah A. Smith, The Risk of Racial Bias in Hate Speech Detection, 2019 Proc. 57th Ann. Meeting Ass'n Computational Linguistics 1668, 1668, https://homes.cs.washington.edu/~msap/pdfs/sap2019risk.pdf [https://perma.cc/YBB9-52D6]; Alessandra Gomes, Dennys Antonialli & Thiago Dias Oliva, Drag Queens and Artificial Intelligence: Should Computers Decide What Is 'Toxic' on the Internet?, Internet Lab (June 28, 2019), http://www.internetlab.org.br/en/freedom-of-expression/drag-queens-and-artificial-intelligence-should-computers-decide-what-is-toxic-on-the-internet [https://perma.cc/CL7X-7AKN].

249. See, e.g., Zuckerberg, Blueprint for Content Governance, supra note 185 ("While I expect this technology to improve significantly, it will never be finished or perfect.").

250. See Mike Ananny, Making Up Political People: How Social Media Create the Ideas, Definitions, and Probabilities of Political Speech, 4 Geo. L. Tech. Rev. 351, 351–56 (2020) [hereinafter Ananny, Making Up Political People].

251. See Jon Porter, WhatsApp Says Its Forwarding Limits Have Cut the Spread of Viral Messages by 70 Percent, Verge (Apr. 27, 2020), https://www.theverge.com/2020/

success in fighting conspiracy theories on the basis that it retrained its recommendation algorithm and reduced watch time of "borderline content" coming from recommendations by seventy percent.[252] Interstitials—like warning screens, fact-check labels or prompts to read articles before sharing them—are measured in relative, not absolute terms, such as how many people clicked through or shared content.[253] Probability is a governing logic across almost all platform design and moderation choices, not merely content takedowns.

Crucially, algorithmic moderation still involves human decisions about speech rights. AI tools are not neutral.[254] Responding to Wu's question, *Will Artificial Intelligence Eat the Law?* due to its increased use in content moderation,[255] Olivier Sylvain argues that "[s]creening technologies do not make content distribution decisions. As rule-bound as they are, they do not eat anything they are told to leave alone."[256] Humans design content moderation systems.[257] Humans tell AI tools what to look for. Humans remain in control of online speech rights. (Or, at least, they should.)

But system-design decisions involved in training and deploying AI do differ in important ways to traditional decisions about speech. Because the decisions are taken ex ante, at the moment of system design and tool development, they are intrinsically *systemic* rather than *individualistic*. They are not made on a case-by-case basis, but by considering the platform as a whole. This makes platform rights adjudication "actuarial."[258] As Professor Mike Ananny explains:

4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline (on file with the *Columbia Law Review*).

252. See Clive Thompson, YouTube's Plot to Silence Conspiracy Theories, WIRED (Sept. 18, 2020), https://www.wired.com/story/youtube-algorithm-silence-conspiracy-theories [https://perma.cc/QR7Q-REBM].

253. See, e.g., Twitter Comms (@TwitterComms), Twitter (Sept. 24, 2020), https://twitter.com/TwitterComms/status/1309178716988354561 (on file with the *Columbia Law Review*) (citing the percentage rates at which people read or shared articles after being prompted to read before sharing).

254. See Langdon Winner, Do Artifacts Have Politics?, 109 Daedalus 121, 134 (1980) ("[I]ntractable properties of certain kinds of technology are strongly, perhaps unavoidably, linked to particular institutionalized patterns of power and authority.").

255. Wu, Will Artificial Intelligence Eat the Law, supra note 22.

256. Sylvain, supra note 54, at 269.

257. See Aziz Z. Huq, A Right to a Human Decision, 106 Va. L. Rev. 611, 646–49 (2020) [hereinafter Huq, A Right to a Human Decision] (describing how machine learning is permeated with normative choices a human designer must select).

258. Frederick F. Schauer, Profiles, Probabilities, and Stereotypes 6 (2009) [hereinafter Schauer, Profiles]; Mike Ananny, Probably Speech, Maybe Free: Toward a Probabilistic Understanding of Online Expression and Platform Governance, Knight First Amend. Inst. (Aug. 21, 2019), https://knightcolumbia.org/content/probably-speech-maybe-free-toward-a-probabilistic-understanding-of-online-expression-and-platform-governance [https://perma.cc/A9C2-BSG4] [hereinafter Ananny, Probably Speech].

> Platform content moderation is . . . probabilistic. It is a confluence of likelihoods: did an algorithmic filter trigger a computational threshold to block offensive content, did enough users within a particular period of time flag a sufficient amount of content to cause an account to be suspended, and did third-party content moderators evenly apply platforms' content standards?[259]

So while the "enforcement capacity" of AI tools is vast,[260] it remains pervaded by error. As one study commissioned by the European Parliament observed, these technologies use "probabilistic methods; errors cannot be completely avoided. At a given level of technical performance, we can usually reduce the false negatives rate . . . , only if we increase the false positives rate . . . . [W]e can improve sensitivity only by worsening specificity, or equivalently we can improve recall only by reducing precision."[261] The prediction that "[w]hen code is law . . . execution is exquisite, and law can be self-enforcing" (so-called "perfect enforcement")[262] has not arrived. Perfect enforcement is still—and, for the imaginable future, will remain—illusory. *Probable enforcement* is reality.

3. *Probability as Pragmatism.* — Readers may instinctively resist reframing speech rights as probabilistic. It seems at odds with the traditional, individualistic conception of rights, especially *speech* rights which are often considered exceptional and fundamental.

This leads to an obvious objection: If the scale is so unmanageable, why not insist on smaller platforms?[263] This argument has intuitive appeal but, even if practically conceivable,[264] breaking up and capping the size of social media platforms would not change the essential problem: Whether concentrated on a few platforms or dispersed across multiple laboratories of online governance, the volume of online speech will only continue to increase.[265] This is what makes the internet so vibrant and such a powerful tool for communication. But this volume is also what makes it so unwieldy and requiring of new thinking about speech rights.

As unsatisfactory as it seems, probabilistic enforcement is the *only* possibility between two extremes of severely limiting speech or letting all

---

259. Ananny, Probably Speech, supra note 258.

260. Wu, Will Artificial Intelligence Eat the Law, supra note 22, at 24.

261. Giovanni Sartor & Andrea Loreggia, Eur. Parliament, The Impact of Algorithms for Online Content Filtering or Moderation: Upload Filters 45 (2020), https://www.europarl.europa.eu/RegData/etudes/STUD/2020/657101/IPOL_STU(2020)657101_EN.pdf [https://perma.cc/94V8-5JLC].

262. Jonathan Zittrain, The Future of the Internet and How to Stop It 107 (2008).

263. For a nuanced version of this argument, see generally Gillespie, Content Moderation, supra note 221.

264. I share Ananny's skepticism on this count. See Ananny, Making Up Political People, supra note 250, at 15.

265. Structural reforms are likely a necessary component of a broader reform agenda. Nothing stated here should be taken to suggest that the proposed reframing of content moderation would be a sufficient or exclusive fix to all problems with platforms.

the posts flow. Some might be attracted to one of these two extremes, but neither is an optimal rights equilibrium. The internet has enabled more broadcasting of expression and amplification of speech, including by those historically marginalized and excluded, than any time in history.[266] Platforms are powerful venues for free expression and a world without them would be a loss for free speech. As argued above,[267] however, the posts-as-trumps approach is no longer tenable. Probabilistic enforcement is an imperfect but pragmatic compromise. Simply put, "Algorithmic moderation is here to stay, now mandated either implicitly or explicitly in both legislation and information platform regulation."[268]

This may not be as radical a departure from traditional thinking about speech as it initially seems. No legal system guarantees a right to an accurate decision.[269] Enforcement of speech rules has never been perfect, online or off. Governance systems always make choices that acknowledge the inevitability of error: Resource allocation, burdens of proof, and due process requirements are just some of the many determinants of error distribution.

Even in the context of speech rights, law accepts that error is inevitable. But the crucial tension is that First Amendment doctrine made a very clear choice about error preference and chose to err all in one direction. As Professor Fred Schauer notes, the "logical and necessary mandate of the chilling effect doctrine," for example, is to "allocate the risk of error away from" chilling speech over chilling other activities.[270] Or, as Professor John Hart Ely put it, First Amendment jurisprudence as a whole comes "to much the same thing: that false-positives are not to be tolerated."[271] Professor Adrian Vermeule, drawing on Professor Vincent Blasi's influential work, similarly observes that much free speech doctrine can be summarized as a kind of "constitutional risk aversion."[272] But Vermeule also called, as this Article does now, for a more mature approach to speech governance that instead takes into account "countervailing risks and harms on all sides of speech protections and speech restrictions."[273]

---

266. See, e.g., Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest 6 (2017) ("Thanks to digital technologies, ordinary people have new means of broadcasting—the potential to reach millions of people at once . . . [and] have greatly increased the ability of citizens to document wrongdoings and potentially move the conversation beyond 'authorities said, activists claimed.'").

267. See supra section I.A.1.

268. Gorwa et al., supra note 223, at 10.

269. See Huq, A Right to a Human Decision, supra note 257, at 653 ("[E]ven in high-stakes contexts such as criminal cases or post-conviction review of capital punishment, the Supreme Court has shied away from a personal right to a *true* determination.").

270. Frederick Schauer, Fear, Risk and the First Amendment: Unraveling the Chilling Effect, 58 B.U. L. Rev. 685, 705 (1978) [hereinafter Schauer, Fear, Risk and the First Amendment].

271. Ely, supra note 78, at 105.

272. Adrian Vermeule, The Constitution of Risk 41 (2013).

273. Id. at 62.

* * *

Recognizing the role of proportionality and probability in online speech governance shows that content moderation is a task of *systemic balancing*: Interests are balanced and error rates are rationalized at the level of system design. Before returning to this in section III.B, this Article examines a case study that vividly illustrates this centrality of proportionality and probability to current online speech governance: content moderation during the COVID-19 pandemic.

D. *The Pandemic*

In some ways, content moderation during the COVID-19 pandemic has been exceptional—an invocation of "emergency powers"[274] and an abandonment of platforms' "defensive crouch" about the extent of their power over online speech.[275] In important ways, however, "[c]ontent moderation during this pandemic is [simply] an exaggerated version of content moderation all the time."[276] Platforms' highly visible interventions have only made more apparent fundamental truths about online speech governance: It involves balancing interests and making judgments about proportionate restrictions on speech when writing rules, and probabilistic choices about error preference when enforcing them.

1. *The Pandemic and Proportionality.* — Platforms have taken an unusually aggressive approach to removing false content during the pandemic, invoking greater risks of harm as justifying more extensive speech restrictions.

Facebook has said that the expanding list of types of content it will disallow is "an extension of our existing policies to remove content that could cause physical harm."[277] Its pandemic policies therefore represent a different balancing of speech and safety interests in the context of the public health emergency. As Facebook's head of global policy management said, "It's always a balance . . . . The approach hasn't changed . . . . [T]his is just a bigger and longer threat."[278]

---

274. Douek, The Internet's Titans, supra note 13.

275. Jack Goldsmith & Andrew Keane Woods, Internet Speech Will Never Go Back to Normal, Atlantic (Apr. 25, 2020), https://www.theatlantic.com/ideas/archive/2020/04/what-covid-revealed-about-internet/610549 (on file with the *Columbia Law Review*) (last updated Apr. 27, 2020).

276. This section draws on Douek, COVID-19 and Social Media Content Moderation, supra note 39.

277. Kang-Xing Jin, Keeping People Safe and Informed About the Coronavirus, Facebook: Newsroom (Jan. 20, 2020), https://about.fb.com/news/2020/03/coronavirus [https://perma.cc/4A8Y-2PW4] (last updated Oct. 5, 2020).

278. Jacob Mchangama, Special Edition—Monika Bickert, Stitcher: Clear & Present Danger, at 03:32–11:10 (Apr. 17, 2020), https://www.stitcher.com/podcast/jacob-mchangama/clear-and-present-danger-a-history-of-free-speech/e/68893679 [https://perma.cc/DU46-BPUB].

Twitter similarly announced that it is "[b]roadening [its] definition of harm to address content that goes directly against guidance from authoritative sources of global and local public health information."[279] Ordinarily, Twitter does not remove content purely because it is false, but the list of pandemic-related misinformation it is removing is long. In a watershed moment, Twitter even removed tweets of world leaders that contained misinformation about COVID-19 cures and added labels to tweets from President Trump.[280] Even before the pandemic, however, Twitter had been developing a gradually expanding set of policies against which the new COVID policies do not look out of place.[281]

The arc of Google's policies has been similar, although harder to pin down given that the company tends to be more opaque. Google's initial pandemic policies were vague, simply stating that it was "removing COVID-19 misinformation" and taking down "thousands of videos related to dangerous or misleading coronavirus information."[282] But YouTube belatedly revealed a more detailed and extensive policy.[283] These were additional to, but not inconsistent with, preexisting policies about harmful content.[284]

These moves reflect the broad trend across the industry. For example, Pinterest started limiting all search results about the coronavirus to results from "internationally-recognized health organizations."[285] Reddit "quarantined" (its term for imposing measures that require users to specifically opt in to see certain content) two subreddits so that users would not view misinformation accidentally.[286] Medium started aggressively taking down

279. Vijaya Gadde & Matt Derella, An Update on Our Continuity Strategy During COVID-19, Twitter Blog (Mar. 16, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html [https://perma.cc/8AAE-37JQ] (last updated Apr. 1, 2020).

280. Evelyn Douek, Trump Is a Problem That Twitter Cannot Fix, Atlantic (May 27, 2020), https://www.theatlantic.com/ideas/archive/2020/05/twitter-cant-change-who-the-president-is/612133 (on file with the *Columbia Law Review*).

281. See Twitter, Twitter Rules, supra note 154.

282. Sundar Pichai, COVID-19: How We're Continuing to Help, Google (Mar. 15, 2020), https://blog.google/inside-google/company-announcements/covid-19-how-were-continuing-to-help [https://perma.cc/92S2-4HCZ].

283. COVID-19 Medical Misinformation Policy, YouTube Help (May 20, 2020), https://support.google.com/youtube/answer/9891785 [https://perma.cc/WRD9-F6XY].

284. See Buni & Chemaly, supra note 84.

285. Craig Silverman, Pinterest Is Blocking Coronavirus Searches, and People Are Very Happy About It, BuzzFeed News (Mar. 13, 2020), https://www.buzzfeednews.com/article/craigsilverman/pinterest-is-blocking-coronavirus-searches-and-people-are [https://perma.cc/7NE2-FEQB].

286. Emma Betuel, Meet the Mods Behind the Fastest-Growing Subreddit: r/Coronavirus, Inverse (Mar. 14, 2020), https://www.inverse.com/mind-body/the-mods-behind-the-fastest-growing-covid-19-subreddit [https://perma.cc/Q24T-X728].

viral posts under a new policy on COVID-19 content, despite the site's mission to be a platform for "whatever you have to say."[287]

This flood of new policies was in one sense akin to a declaration of emergency, with platforms mobilizing to impose guardrails on the internet in the face of the "infodemic."[288] From a broader perspective, however, these policies were simply reflective of the growing trend of platforms balancing speech rights against other interests, albeit with the recognition that in the context of a global pandemic these other interests are particularly weighty.

2. *The Pandemic and Probabilities*. — Just as platforms were implementing these more aggressive policies, they were faced with the same challenge confronting businesses around the world: Their content moderation contractors and staff had to work from home. For logistical reasons, platforms drastically scaled back human moderation and increased their reliance on AI.[289]

In a break from the usual optimism about the efficacy of such tools,[290] platforms acknowledged that this shift to AI-centric moderation would result in more errors. Facebook,[291] Twitter,[292] and YouTube[293] all admitted that relying on AI tools without human moderation would result in blunter, less contextualized decisions and more mistakes.

Generally, these announcements presaged more "false positives"—that is, over-removal of content. Early reports from platforms of the consequences of these choices confirmed that this was the general effect.[294]

---

287. COVID-19 Content Policy, Medium Help Ctr., http://help.medium.com/hc/en-us/articles/360045484653 (on file with the *Columbia Law Review*) (last visited Nov. 16, 2020); Ev Williams, Welcome to Medium, Medium (Aug. 14, 2012), https://medium.com/@ev/welcome-to-medium-9e53ca408c48 [https://perma.cc/V88W-XL4B].

288. Douek, COVID-19 and Social Media Content Moderation, supra note 39.

289. Id.

290. See, e.g., Harwell, supra note 227 (describing Mark Zuckerberg's insistence that "[a]rtifical intelligence will solve Facebook's most vexing problems").

291. Jin, supra note 277 ("[W]e will now rely more on our automated systems to detect and remove violating content and disable accounts. As a result, we expect to make more mistakes . . . .").

292. Gadde & Derella, supra note 279 (stating that Twitter would increase its use of machine learning and automation, and that these systems "sometimes lack the context that our teams bring, and this may result in us making mistakes").

293. Protecting Our Extended Workforce and the Community, YouTube Off. Blog (Mar. 16, 2020), https://blog.youtube/news-and-events/protecting-our-extended-workforce-and [https://perma.cc/W3VC-EXWF] ("[A]utomated systems will start removing some content without human review . . . . As we do this, users and creators may see increased video removals, including some videos that may not violate policies.").

294. See, e.g., Alex Barker & Hannah Murphy, YouTube Reverts to Human Moderators in Fight Against Misinformation, Fin. Times (Sept. 20, 2020), https://www.ft.com/content/e54737c5-8488-4e66-b087-d1ad426ac9fa (on file with the *Columbia Law Review*) (describing the drastic increase in Facebook's removal of hate speech content); Issie Lapowsky, How COVID-19 Helped—and Hurt—Facebook's Fight Against Bad Content, Protocol (Aug. 11,

Although platforms could not help sending their human moderators home, another possible error choice was available: reducing use of AI tools and allowing more false negatives. That is, they could have moderated far less in general and tolerated more violating content. But in the context of the "infodemic" and global emergency, this was not much of an alternative at all. Platforms would quickly become unusable, mired in spam and—during the pandemic especially—actively harmful misinformation.[295] The choice between false positives and false negatives was therefore an unusually easy one. But less extreme versions of this choice happen in the context of *every* content moderation policy.

The hyper-reliance on AI tools during the pandemic was merely an unusually visible, honest, and exaggerated example of the way content moderation always works. As Stamos has explained, policy teams work hand-in-hand with tech and operational teams within the company, and policies are designed based on comparing false positive rates across tests of different rules.[296] Error rates are not merely a consequence of policy choices; they are deeply embedded in them.

Content moderation during COVID-19 therefore represented a dramatic rhetorical shift from platforms, but not a truly substantive one. In some ways, it simply represented an unusual moment of candor about the governing logics of content moderation system design.

But if the pandemic is a clarifying example of the inherent dynamics of content moderation, it is also a stark illustration of its ongoing deficits. Despite early acclaim,[297] platforms' efforts to stem the tide of harmful content quickly ran into criticism as it became clear that even—or perhaps especially—under emergency conditions, drawing lines can be hard and controversial. This was compounded by a lack of transparency that led to confusion and distrust about exactly what interests were being balanced. Why could some world leaders suggest unproven remedies for COVID-19 while others had their posts removed for the same offense?[298] Why were

---

2020), https://www.protocol.com/covid-facebook-content-moderation [https://perma.cc/BRE3-QHBU].

295. Sarah T. Roberts, Over*Flow: Digital Humanity: Social Media Content Moderation and the Global Tech Workforce in the COVID-19 Era, Flow J. (Mar. 19, 2020), https://www.flowjournal.org/2020/03/digital-humanity [https://perma.cc/9DUJ-X7SL].

296. Mathew Ingram & Alex Stamos, Alex Stamos Talks About Facebook's Oversight Board, Galley by CJR, https://galley.cjr.org/public/conversations/-M74eLMfvkdKpIPjRfo4 (on file with the *Columbia Law Review*) (last visited Oct. 23, 2020).

297. See Casey Newton, How COVID-19 Is Changing Public Perception of Big Tech Companies, Verge (Mar. 26, 2020), https://www.theverge.com/interface/2020/3/26/21193902/tech-backlash-covid-19-coronavirus-google-facebook-amazon (on file with the *Columbia Law Review*); Ben Smith, When Facebook Is More Trustworthy than the President, N.Y. Times (Mar. 15, 2020), https://www.nytimes.com/2020/03/15/business/media/coronavirus-facebook-twitter-social-media.html (on file with the *Columbia Law Review*).

298. Kim Lyons, Twitter Removes Tweets by Brazil, Venezuela Presidents for Violating COVID-19 Content Rules, Verge (Mar. 30, 2020), https://www.theverge.com/2020/3/30/21199845/twitter-tweets-brazil-venezuela-presidents-covid-19-coronavirus-jair-

some anti-lockdown protest events allowed while others were removed?[299] Furthermore, ongoing failures to fully explain error choices led to continued frustration and misunderstanding as the mistakes that platforms had forewarned about, and which were accepted in theory, came to fruition and seemed less acceptable in practice. The over- and under-enforcement of the mask-ad ban example with which this Article opens is a case in point.[300]

The next Part describes these challenges in more detail. Just as it highlighted the centrality of proportionality and probability, content moderation during the pandemic also highlighted the lack of transparency and vocabulary necessary to intelligently evaluate the trade-offs inherent in putting these principles into practice.

II. ASKING THE RIGHTS QUESTIONS

So far, this Article has shown that proportionality and probability are the precepts around which online speech governance is now crafted, and it has defended this new paradigm of systemic balancing. But these shifts require a reconceptualization of the online speech governance project. Yet, because these changes have generally occurred incrementally and implicitly, they remain undertheorized and discourse around them impoverished.

This Part turns to the weaknesses and questions that have been neglected within this new framework, and which undermine the actual and perceived legitimacy of current governance. There are two central unaddressed weaknesses: (1) a lack of clarity about what interests should be taken into account in the balancing exercise and how to deal with their "incommensurability"; and (2) the need to openly acknowledge the inevitability of error and the error choices this necessitates.

A. *Incommensurability*

As soon as the decisionmaker moves from professing to be a mere "taxonomist" under a categorical approach to a "grocer" weighing competing values, questions necessarily arise about what gets measured and how the scales are calibrated. It is the very subjectivity, and some would say impossibility, of this exercise that has been the basis of American

---

bolsonaro-maduro (on file with the *Columbia Law Review*); Andrew Solender, All the Times Trump Had Promoted Hydroxychloroquine, Forbes (May 22, 2020), https://www.forbes.com/sites/andrewsolender/2020/05/22/all-the-times-trump-promoted-hydroxychloroquine/?sh=17adc0694643 [https://perma.cc/HLC8-V74B].

299. Sam Adler-Bell, Facebook Is Removing Protest Pages. That's a Terrible Precedent., Medium: OneZero (Apr. 24, 2020), https://onezero.medium.com/facebook-is-removing-protest-pages-thats-a-terrible-precedent-a2fabd904f63 [https://perma.cc/ZAZ6-UB4C].

300. See supra notes 4–10 and accompanying text.

resistance to balancing in constitutional law.[301] This is most memorably encapsulated in Justice Scalia's declaration that balancing is doomed to fail because competing interests are incommensurate and so trying to balance them is "like judging whether a particular line is longer than a particular rock is heavy."[302] The incommensurability debate has felled many trees, and I shall not settle it here. My own view is that the indeterminacy critique is overstated and, perhaps more importantly, the determinacy of any alternative is also overstated.[303] As Professor Richard Fallon argues, "Along a myriad of often unrecognized dimensions, constitutional law requires the identification, specification, weighing, balancing, and accommodation of sometimes competing individual and governmental interests."[304] There is no escaping this task.[305]

But the more fundamental point for present purposes is that this problem is not new. The evolution of content moderation debates mirrors fundamental debates about rights adjudication generally, the exceptionalism of the American categorical approach to free speech jurisprudence, and broader arguments of rules versus standards.[306]

But in the context of content moderation, general concerns about balancing take an even sharper form. Not only do private platforms lack any constitutional mandate or legitimacy to conduct the fraught balancing exercise, it is not even clear what that exercise encompasses in such a context. Employing proportionality as a method of analysis cannot answer the anterior question of *what* needs to be proportionate to *what*. Proportionality is a tool—a "framework that must be filled with content."[307] Two questions need to be answered: What interests should be taken into account, and how do you determine their relative weights? Neither question can be directly translated from a state-based system of adjudication.

The first question—what purposes are "legitimate" reasons for limiting speech—is distinctive in the context of content moderation because of the private nature of platforms. In constructing and enforcing online speech rules, platforms are not neutral parties, but businesses with

---

301. See Sullivan, supra note 17, at 293–94 ("[Justices Scalia and Kennedy] have condemned balancing for affording judges excessive discretion.").

302. Bendix Autolite Corp. v. Midwesco Enters., 486 U.S. 888, 897 (1988) (Scalia, J., concurring).

303. See Douek, All Out of Proportion, supra note 168, at 567–70 (arguing that a categorical approach to constitutional law does not necessarily reduce uncertainty).

304. Richard H. Fallon, Jr., The Nature of Constitutional Rights: The Invention and Logic of Strict Judicial Scrutiny 7 (2019) [hereinafter Fallon, Nature of Constitutional Rights].

305. See Stephen Breyer, The Court and the World: American Law and the New Global Realities 257 (2015) ("[E]ven judges who explicitly write only in terms of categories are implicitly balancing harms and objectives . . . .").

306. See Schauer, Exceptional First Amendment, supra note 71, at 30–32.

307. Aharon Barak, Proportionality (2), *in* Oxford Handbook of Comparative Constitutional Law 738, 741 (Michel Rosenfeld & András Sajó eds., 2012).

a stake in aggregate outcomes (if not individual ones).[308] Content moderation is *the* commodity platforms offer,[309] and they have their own First Amendment rights (or, in international parlance, "speech rights") to determine what content to host and how they present it.[310] Indeed, they have a fiduciary duty to maximize stockholder value.[311] The extent to which these interests can be weighed against (or in favor of) individual speech rights and other social interests is essentially open.[312]

Existing bodies of law do not have preexisting tools to account for these interests and the different balancing calculus they create.[313] Facebook, for example, has adopted its own "values" that can justify limitations on speech which have some overlap, but are not coextensive, with those in other legal systems.[314] How should any conflicts between Facebook's values and those recognized by state-based free speech doctrines be reconciled? Perhaps the simplest answer is that state-based norms can provide a floor below which private platform interests cannot justify going. But this answers only the obvious cases, such as not allowing a platform that professes to be open to political debate to restrict such debate on the basis of viewpoint alone. But most of the contentious issues of content moderation happen in the vast space where there are multiple possible rights-respecting alternative answers.

Transplanting from state-based jurisprudence also does not account for other characteristics that make platform content moderation different from state action: Most obviously, while there can be serious consequences from having content or accounts removed from social media,[315] these will

---

308. See Van Loo, supra note 40 (manuscript at 27).

309. See Gillespie, Custodians of the Internet, supra note 59, at 5 (arguing that platforms have been forced to moderate content in order to maintain their continued existence by attracting advertisers and users).

310. Eric Goldman, Of Course the First Amendment Protects Google and Facebook (and It's Not a Close Question), Knight First Amend. Inst. (Feb. 26, 2020), https://knightcolumbia.org/content/course-first-amendment-protects-google-and-facebook-and-its-not-close-question [https://perma.cc/D6X4-QGUP].

311. Lina M. Khan & David E. Pozen, A Skeptical View of Information Fiduciaries, 133 Harv. L. Rev. 497, 503–04 (2019).

312. Molly K. Land, Regulating Private Harms Online: Content Regulation Under Human Rights Law, *in* Human Rights in the Age of Platforms, supra note 55, at 285, 292–93.

313. Id. at 290 ("There are several different types of human rights problems raised by the processes of content moderation, curation, and account suspension/termination. Nonetheless, the extent to which human rights law governs these activities is unclear."); see also Emily B. Laidlaw, Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility 243 (2015); Barrie Sander, Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, 43 Fordham Int'l L.J. 939, 969–70 (2020).

314. Douek, Why Facebook's Update Matters, supra note 28.

315. See Hill, supra note 199 (recounting the exasperation of some users kicked off Facebook, and noting that one Facebook user whose account was disabled, for example, felt

usually fall short of the consequences of state sanction. Conversely, most algorithmic moderation occurs at upload, and therefore operates as a prior restraint.[316] State-based jurisprudence does not speak to how restriction of speech by a private platform is simply just *different* to traditional government restrictions.

Making matters worse still, even if it were clear which interests should be taken into account, platforms have no particular competency in assessing them.[317] As U.N. Special Rapporteur David Kaye noted, "Companies are not in the position of Governments to assess threats to national security and public order, and hate speech restrictions on those grounds should be based not on company assessment but legal orders from States."[318] This might imply that companies *are* in a good position to evaluate restrictions on speech for other purposes such as rights of others, public health, or morals.[319] But it is not clear that is the case, especially for quintessentially Silicon Valley–based companies purporting to determine public health or morals for users in, say, India[320] or Europe.[321]

If under the posts-as-trumps approach platforms half-innocently asked "who are *we* to balance rights and interests?"[322] when platforms conduct proportionality analysis, others now ask the same thing: "Who are *you* to decide the balance between competing interests?"[323] Platforms are making consequential decisions about the shape of public discourse. They

---

as if Facebook was "holding [the user's] social network hostage, along with [their] memories").

316. See Jack M. Balkin, Old-School/New-School Speech Regulation, 127 Harv. L. Rev. 2296, 2318 (2014).

317. Giancarlo Frosio, Mapping Online Intermediary Liability, *in* Online Intermediary Liability, supra note 235, at 1, 26 ("Online intermediaries are unequipped—and lack constitutional standing—for making decisions involving a proportional balancing of rights.").

318. Kaye, Freedom of Opinion and Expression, supra note 184, at 18.

319. These are the other legitimate grounds for limitation of a right under the ICCPR, supra note 203, art. 19.

320. Olivia Solon, Inside Facebook's Efforts to Stop Revenge Porn Before It Spreads, NBC News (Nov. 18, 2019), https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631 [https://perma.cc/K9LA-LPF6] (last updated Nov. 19, 2019) ("[A] woman in India . . . reported a photo in which she was fully clothed in a pool with a fully clothed man. 'Within her culture and family that would not be acceptable behavior,' [Facebook representative Antigone] Davis said. 'The photo was shared intentionally . . . to harass her.'").

321. Chris Marsden, Trisha Meyer & Ian Brown, Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?, Comput. L. & Sec. Rev., Apr. 2020, at 1, 17 ("Executives in California, ex-politicians such as Nick Clegg, or thousands of badly-paid contractors hired off the internet, from the Philippines or India, cannot regulate European fake news: it has to be Europeans.").

322. See supra section I.A.3.

323. See, e.g., Marco Bassini, Fundamental Rights and Private Enforcement in the Digital Age, 25 Eur. L.J. 182, 191 (2019) ("[T]he fact that the decision to evaluate the existence of a public interest to access information is left to private actors is concerning from a substantive perspective.").

might base these choices on balancing public interests or their own private interests, but in neither case are they well-placed to be trusted arbiters. They have neither doctrine nor competency to bring to bear on the determination of how these interests should be weighted. In attempting to apply proportionality in content moderation, not only are platforms comparing the length of a line with the weight of a rock, but they must do it using only a hammer while wearing a blindfold.

Nevertheless, the only way out is through. There is no avoiding the fundamental task of balancing interests in writing content moderation rules. Governments can set parameters within the limits of their constitutions, but they will never be able to answer most online speech governance questions, both because of the speed and scale at which such governance must happen and because governmental involvement in the minutiae of speech control is inherently suspect and generally unconstitutional. Therefore, the question is not *if* but *how* platforms should perform this impossible task. Part III below expands on this, but first, things get more dire still. The next section turns to the problems that probability raises for speech governance.

B. *Getting Rights Wrong*

Content moderation systems must accept error. Indeed, they decide to get it wrong sometimes and must decide in which direction to do so. The choice to get it wrong in some portion of cases is the price of getting it right, within a reasonable timeframe (or at all), in the vast majority of cases. Therefore, the question a probability framework demands an answer to is: How should decisionmakers compare different kinds of errors and error rates?

Academics, lawmakers, and the broader public discourse have almost entirely eschewed this kind of analysis.[324] Instead, frequent public outrage cycles focus on *instances* of platforms failing to perfectly enforce their rules. Typical of the genre are the stories about Facebook's failure to properly enforce its mask-ad ban from the introduction.[325] Another example is Facebook deplatforming antiracist skinheads while attempting to remove white supremacy groups.[326] There is no denying, and indeed Facebook acknowledged, that these were mistakes.[327] But when error is inevitable,

---

324. An exception is conversations in the copyright context. See, e.g., Ben Depoorter & Robert Kirk Walker, Copyright False Positives, 89 Notre Dame L. Rev. 319, 328 (2013) (arguing for greater consideration of the cost of false positives in copyright enforcement).

325. Isaac, supra note 5.

326. Chloe Hadavas, Why We Should Care that Facebook Accidentally Deplatformed Hundreds of Users, Slate (June 12, 2020), https://slate.com/technology/2020/06/facebook-anti-racist-skinheads.html [https://perma.cc/4Z4E-XDGA].

327. Sarah Emerson, Facebook Deplatforms Hundreds of Anti-Racist Skinheads and Musicians, Medium: OneZero (June 10, 2020), https://onezero.medium.com/facebook-deplatforms-hundreds-of-anti-racist-skinheads-and-musicians-6b57bef204e1 [https://perma.cc/W5B4-7YVJ] ("'We apologize to those affected by this issue,' a Facebook spokesperson

these stories need reframing. The question cannot be whether there are *instances* of false positives (benign mask content being taken down) or false negatives (mask ads continuing to run) when enforcing a mask-ad ban—the answer to that question will *always* be "yes." The more pertinent questions are: What error rates are acceptable in enforcing a ban on mask ads? Should Facebook err on the side of removing some volunteer mask makers for the benefit of quickly removing most mask ads (over-enforcement), or on the side of trying to ensure no volunteer mask-makers get caught but some mask ads go undetected (under-enforcement)? This answer in part depends on technical capacity to calibrate between these two extremes.

This reframing might cause discomfort for two reasons in particular. First, there is a risk of allowing error rates to become excuses for negligent enforcement practices and inadequate investment by platforms in content moderation. The second is the reason described earlier,[328] of error analysis being at odds with traditional thinking about speech rights, which are individualistic and, to the extent they consider systemic error costs, always err on the side of avoiding false positives.

The first risk is what Professor Sarah Roberts calls content moderation's "logic of opacity":[329] the cultivated secrecy that allows companies to frame individual and high-profile cases of error as reasonable mistakes while preventing interrogation of the systems that led to them. A stark example of this is Zuckerberg's initial explanation of a failure to remove a militia page calling for people to bring weapons to a protest in Kenosha, Wisconsin, as an "operational mistake," despite later revelations that an event created by the page had been reported 455 times, making up sixty-six percent of all event reports that day.[330] In the abstract, failing to take down a single event page can seem like an individual tragic error; but with the relevant contextual information, ignoring such a large and disproportionate number of reports cannot be seen as anything other than a systemic failure. Thus, crucially, accepting that errors are inevitable must be the beginning of the conversation, not the end of it. Determining acceptable error rates and which errors platforms should err on the side

---

told *OneZero* following the publication of our report. 'These accounts were removed in error and have been reinstated.'"); Isaac, supra note 5 ("'The automated systems we set up to prevent the sale of medical masks needed by health workers have inadvertently blocked some efforts to donate supplies,' Facebook said in a statement. 'We apologize for this error . . . .'"); Silverman, Facebook Mask Ad-Ban, supra note 6 ("After an investigation, the company said it has banned ZestAds from its platform.").

328. See supra section I.C.

329. Sarah T. Roberts, Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation, First Monday (Mar. 5, 2018), https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649 [https://perma.cc/ZH4K-J9WY].

330. Ryan Mac, A Kenosha Militia Facebook Event Asking Attendees to Bring Weapons Was Reported 455 Times. Moderators Said It Didn't Violate Any Rules., BuzzFeed News (Aug. 28, 2020), https://www.buzzfeednews.com/article/ryanmac/kenosha-militia-facebook-reported-455-times-moderators [https://perma.cc/MX4E-N735].

of making is impossible in the absence of adequate information about the systems that make and enforce those decisions and their capabilities and biases. But this too can be a strength of focusing on systemic rather than individual choices: Systemic values and problems can only be identified from this broader perspective.[331]

This reflects discourse about systemic biases and governance of AI more generally.[332] Content moderation discourse has been slow to join this conversation because of the strong hold that the individual framing of speech rights has on our imagination. If under the posts-as-trumps framework, the platforms asked "who are *we* to balance rights and interests?" and the proportionality framework has society asking "who are *platforms* to balance rights and interests?", then the move to probability raises the question: "Who is *anyone* to plan to get decisions about free speech wrong?"

But online speech governance needs to learn to live with this discomfort. The internet has unleashed a torrent of expression that needs to be governed. Whether it be on platforms currently dominant or new and smaller ones, there is no escaping that the traditional individualistic understanding of speech rights does not compute in these new systems of free expression.

Of course, the individualistic conception is itself somewhat of a myth. Legal systems always make errors. The problem is that historically they have not been so legible, in the sense that every instance of erroneously categorized content is recorded and often searchable. Nor have legal systems historically been so brazen or candid about their mistakes. Error choice—especially in rights cases—is often viewed with embarrassment. Schauer argues this is unfortunate:

> Defending the making of mistakes is a formidable task. It may even be unwise. Yet . . . creating a decisionmaking procedure predicted at the outset to make some number of errors will often lead to fewer errors in the long run than will creating a decisionmaking procedure that in theory produces no errors but that in practice produces many.[333]

If academic and public discourse around content moderation has so far largely failed to engage in this kind of analysis, it is not because it is impossible: Platforms do it *internally* all the time.[334] Policy teams within platforms work "hand-in-hand" with tech and operational teams so that any rule changes can be tested on random samples to "see how effective

---

331. See Aziz Z. Huq, Constitutional Rights in the Machine Learning State, 105 Cornell L. Rev. 1875, 1937 (2020).

332. See Bloch-Wehba, Automation in Moderation, supra note 45, at 47 n.250, and sources cited therein.

333. Schauer, Profiles, supra note 258, at 23.

334. See, e.g., Facebook's Civil Rights Audit, supra note 25, at 47 (reporting that Facebook "studies the accuracy of content decisions and seeks to identify the underlying causes of reviewer error").

the rules are, how many false positive[s] will exist, and whether there are any unforeseen consequences."[335] Ex-Facebook Chief Security Officer Stamos gave the example of Facebook's expansion of its ban on dangerous groups to include those espousing "white nationalism" as a policy change that would be subjected to this kind of testing for how easy it would be to operationalize.[336] As Facebook later described the decision, writing the policy took nearly a year, largely because the company was working to determine "what are the kinds of speech that might actually unintentionally be swept in to this policy . . . depending on how we draw the line."[337] As this example shows, even decisions that seem like quintessentially questions of principle (are white nationalist groups dangerous organizations?) are subject to probabilistic enforcement calculations.

External discourse, by contrast, largely focuses on matters of free speech principle and theory rather than practical enforcement constraints. Neither is irrelevant: Being able to justify policies as a matter of principle remains important, but *only* justifying policies on the basis of principle regardless of how they operate in practice is inadequate. But enforcement constraints also cannot obscure the important free speech values at stake. Probabilistic pragmatism is not mistaken but unavoidable; to legitimize it, error choice calibration cannot remain entirely hidden from view.

A rare public example of this kind of analysis comes from cloud service provider Cloudflare, which made its child sexual abuse material (CSAM) detection technology available to all its clients.[338] Its announcement of the move provided rare insight into the trade-offs platforms make. CSAM detection technology works by checking uploads to a platform against a central repository of CSAM images.[339] Cloudflare noted that one of the biggest questions was "what the appropriate threshold should be" for matches.[340] Too strict a threshold and there will be many false negatives (CSAM left unflagged), but too loose and there will be many false positives. The post explained that while "[f]alse positives may seem like the lesser evil . . . there are legitimate concerns that increasing the possibility of false positives at scale could waste limited resources and further overwhelm the

335. Ingram & Stamos, supra note 296.

336. Id.

337. Press Release, Facebook, Press Call on Sixth Edition of Community Standard Enforcement Report 9 (Aug. 11, 2020), https://about.fb.com/wp-content/uploads/2020/08/Press-Call-Transcript.pdf [https://perma.cc/DR69-Y9DW].

338. Justin Paine & John Graham-Cumming, Announcing the CSAM Scanning Tool, Free for All Cloudflare Customers, Cloudflare Blog (Dec. 18, 2019), https://blog.cloudflare.com/the-csam-scanning-tool [https://perma.cc/J97B-YEQ6].

339. Evelyn Douek, Knight First Amend. Inst., The Rise of Content Cartels 7–8 (2020), https://s3.amazonaws.com/kfai-documents/documents/34cfde322f/2.11.2020_Douek_MWFinal.pdf [https://perma.cc/DZF7-EJHL] [hereinafter Douek, Content Cartels].

340. Paine & Graham-Cumming, supra note 338.

existing ecosystem."[341] Different platforms may want different thresholds, depending on their type of business, content review systems, and likely exposure and risk tolerance. Because of all these variables, Cloudflare concluded that it would allow its customers to set their own thresholds: "[A]llowing individual site owners to set the parameters that make the most sense for their particular site will result in lower false negative rates (i.e., more CSAM being flagged) than if we try and set one global standard for every one of our customers."[342]

Thus, even for CSAM, a relatively defined category about which there is essentially universal agreement on the harm it causes and the urgency with which it needs to be removed, placing the decision within operational constraints means that a tighter threshold is not always better. Too many false positives may make identifying and prosecuting actual instances of CSAM more difficult.[343] Or, in cases like the "Terror of War" photo, there are reputational costs for platform mistakes—Facebook accidentally removing this iconic photo of a naked child fleeing a napalm attack during the Vietnam War is still often invoked as an embarrassing footnote in content moderation history.[344] Therefore, even with a relatively clear-cut category, where "[t]he margin of over-removal and collateral damage to lawful speech is likely to be small," there are still no easy answers.[345]

Similar analysis likely takes place for the setting of any rule and the deployment of any content moderation technology.[346] The objective here is not to defend any particular choice of thresholds or error rates,[347] but simply to illustrate that these decisions are choices between errors. Too often, many in these conversations presume a level of technical capability

---

341. Id.

342. Id.

343. See Michael H. Keller & Gabriel J.X. Dance, The Internet Is Overrun with Images of Child Sexual Abuse. What Went Wrong?, N.Y. Times (Sept. 29, 2019), https://www.nytimes.com/interactive/2019/09/28/us/child-sex-abuse.html (on file with the *Columbia Law Review*) ("[G]iven the overwhelming number of reports, it is not uncommon for requests from the authorities to reach companies too late.").

344. Julia Carrie Wong, Mark Zuckerberg Accused of Abusing Power After Facebook Deletes 'Napalm Girl' Post, Guardian (Sept. 9, 2016), https://www.theguardian.com/technology/2016/sep/08/facebook-mark-zuckerberg-napalm-girl-photo-vietnam-war [https://perma.cc/Z2HB-BMS4].

345. Keller, Observations, supra note 225, at 19.

346. See generally Douek, Content Cartels, supra note 339 (discussing the use of similar technology in the context of terrorist and extremist materials, and its likely spread to other contexts); Roshan Sumbaly, Mahalia Miller, Hardik Shah, Yang Xie, Sean Chang Culatana, Tim Khatkevich, Enming Luo, Emanuel Strauss, Gergely Szilvasy, Manika Puri, Pratyusa Manadhata, Benjamin Graham, Matthijs Douze, Zeki Yalniz & Hervé Jegou, Using AI to Detect COVID-19 Misinformation and Exploitative Content, Facebook: AI (May 12, 2020), https://ai.facebook.com/blog/using-ai-to-detect-covid-19-misinformation-and-exploitative-content [https://perma.cc/U4WY-AZMK] (discussing similar trade-offs in the context of visual pandemic misinformation).

347. It is essentially impossible to do so in the absence of proper verified evidence about the actual error rates involved.

that eschews talking about error at all.[348] Lawmakers in particular—no doubt in part because of years of false assurances from platforms—are beginning to enshrine into law short removal deadlines that make reliance on AI tools a practical necessity.[349] But these conversations rarely make explicit the fact that such requirements are themselves an error choice, forcing platforms into over-removal. Instead, such laws proceed on the false assumption that platforms could remove the bad without the good, and faster, if only they just tried harder.[350] Without question, in most cases platforms could likely be doing better, but a simple "nerd harder" directive that does not recognize trade-offs is not sound.

Instead, as Schauer counselled,[351] accepting errors and designing a process that seeks to optimize around them, rather than ignoring them, is likely to be more accurate overall.

To be clear, there may be cases where error costs are too great, or where the probabilities cannot be made to fall within an acceptable range. Accepting errors does not mean accepting *all* errors. But knowing what relevant error rates actually are opens this conversation too, by showing exactly how content moderation alone might be an insufficient tool to tackle a particular problem if reasonable rules are not capable of being enforced within an acceptable range of accuracy. It may be, for example, that the costs of violent livestreams are so high, and the technological difficulty of moderating live footage so difficult, that it outweighs the benefits of such technology being widely available at present.[352] A conversation that assumes you can have the good of livestreams without the bad curtails this conversation entirely.[353]

III. RECALIBRATING

This is the inflection point at which content moderation stands then: Major platforms operate on the basis of systemic balancing, but this is underrecognized and undertheorized. As regulators begin to transform the legal constraints on these systems and platforms emerge from their

348. See Bloch-Wehba, Automation in Moderation, supra note 45, at 48–50.

349. See id. at 33–36.

350. See, e.g., Douek, Nerd Harder, supra note 34 ("The AVM Act exists in a fictional world in which it is possible to have the best of Facebook Live without the worst of it.").

351. See supra note 333 and accompanying text.

352. See Douek, Nerd Harder, supra note 34 (discussing the costs of Facebook Live and its "unfettered availability"); Mary Anne Franks, Justice Beyond Dispute, 131 Harv. L. Rev. 1374, 1377–78 (2018) (reviewing Ethan Katsh & Orna Rabinovich-Einy, Digital Justice: Technology and the Internet of Disputes (2017)) ("Given that live-stream content must still first be flagged by a user before a moderator can review it, the only way that removals can be done more quickly and effectively is for more users to watch and flag more traumatizing content.").

353. See Douek, Nerd Harder, supra note 34, at 27 ("There are important discussions to be had, equities to be balanced, and trade-offs to be made. The AVM Act, by treating the issue as simple, left no room for any of this to take place.").

pandemic "states of emergency," addressing these deficits is crucial. The regulations and institutions that emerge from this moment of disruption will profoundly shape online speech and public discourse.

This Part describes the project of recalibrating content moderation around systemic balancing. First, it addresses concerns that recognizing proportionality and probability will inevitably devalue speech rights. To the contrary, open acknowledgment of these principles can be speech enhancing and protective. Second, this Part describes the questions that recalibrating online speech governance institutional design around systemic balancing raises for platforms and regulators in turn. The conclusions are necessarily tentative: The task requires intellectual humility and incrementalism. Ultimately, online speech governance may converge around certain rules or principles, but the only way to find those principles is through a period of diversity and transparent experimentation. Finally, this Part reiterates the importance of learning these lessons now, as the landscape is unusually destabilized and open for contestation.

A. *Against Speech Rights Deflation*

The arguments in this Article so far may have made many free speech scholars nervous. Previous Parts have talked about proportionality analysis—often seen as anathema to U.S. rights, and especially First Amendment, jurisprudence—and of reducing sacred speech rights to mere probabilities. This subpart hopes to ease those anxieties.

One of the most prominent strands of proportionality critique is that it undermines rights by reducing them to one of many interests to be balanced.[354] This is, after all, the thrust of Dworkin's argument that to subject rights to balancing is to deny them altogether.[355] In the context of speech rights specifically, this is said to be especially pernicious as uncertainty leads to "chilling effects," with those unsure avoiding speaking.[356] If this wasn't bad enough, probabilistic reasoning inherently *accepts* that infringement on valuable speech is inevitable. These two dynamics combined might be thought to inherently devalue and deflate online speech rights. But systemic balancing will not lead to a free speech dystopia for four reasons.

First, this Article's argument has been that online speech governance is already being shaped by proportionality and probability, whether it is explicitly recognized or not. To resist systemic balancing is to ignore reality. The obvious retort is that current content moderation is a free speech

354. See, e.g., Webber, The Cult of Constitutional Rights Scholarship, supra note 168, at 198; Alison L. Young, Proportionality Is Dead: Long Live Proportionality!, *in* Proportionality and the Rule of Law Rights, Justification, Reasoning 43, 44 (Grant Huscroft, Bradley W. Miller & Grégoire Webber eds., 2014).

355. Dworkin, supra note 51, at xi, 192.

356. Schauer, Fear, Risk and the First Amendment, supra note 270, at 693.

dystopia that undervalues speech.[357] But value judgments about what speech is allowed online are inevitable; the question that remains is how to make them. Neither free speech absolutism nor the First Amendment's highly protective regime provides a workable basis for content moderation rules: Platforms would quickly find themselves overrun with spam, adult content, and other merely "unpleasant" content that would diminish the value of their products to users.[358] At the very least, there will always need to be a balance struck between the free speech rights (and business interests) of platforms to decide what content they want to host and the free speech interests of users to say whatever they want. Denying that balancing is taking place in these decisions is only to obscure it from view, not to eliminate it. Exposing such balancing allows for productive argument about it, which could result in its eventual rejection.

This is the second point: Proportionality itself does not demand particular substantive results. Indeed, this is one of the major strands of criticism directed at the method: It is too indeterminate.[359] Accordingly, adoption of proportionality does not inevitably entail a less speech-protective standard than a categorical approach.[360] Of course, recognizing and balancing other interests does suggest more limitations on speech. But just as categorical reasoning can be used to support protecting or censoring certain types of content,[361] it is possible to apply balancing tests more or less robustly and to weight interests differently.[362] The proof of this possible plurality is in the different substantive laws of free expression around the world despite proportionality's global dominance.[363] As Professor Vicki Jackson argues, "[T]here is no conceptual obstacle to providing strong rights protection through proportionality analysis by treating . . . the value of freedom of expression as presumptively stronger than reasons for suppression."[364]

---

357. See Kyle Langvardt, Regulating Online Content Moderation, 106 Geo. L.J. 1353, 1358 (2018).

358. See Gillespie, Custodians of the Internet, supra note 59, at 5.

359. Petersen, supra note 170, at 38–47; see also Frederick Schauer, Balancing, Subsumption, and the Constraining Role of Legal Text, *in* Institutionalized Reason: The Jurisprudence of Robert Alexy 307, 307–08 (Matthias Klatt ed., 2012) [hereinafter Schauer, The Constraining Role]; Schlink, supra note 168, at 299; Webber, The Cult of Constitutional Rights Scholarship, supra note 168, at 196.

360. See Sullivan, supra note 17, at 307–08.

361. See supra section I.A.3.

362. See Julian Rivers, Proportionality and Variable Intensity of Review, 65 Cambridge L.J. 174, 176 (2006).

363. See Barak, Constitutional Rights and Their Limitations, supra note 73, at 489–90 ("Proportionality is a legal framework that must be filled with content. It allows for different levels of protection, according to the principles and values of each legal system.").

364. Jackson, supra note 15, at 3168–69.

Third, because proportionality requires using the "least restrictive means" to limit speech,[365] it can be speech-protective overall. Indeed, looking for a proportionate response will often suggest *more* speech-protective measures. Labeling content as fact-checked or manipulated follows the hallowed counterspeech tradition that the best remedy for falsehoods is more speech, rather than censorship.[366] De-amplification does not reduce the amount of speech and does not directly impede the ability to speak. Whether speech is amplified by platforms' algorithms is a separate question from whether it can be posted in the first place, just as government-subsidized speech is a separate question from regular viewpoint discrimination under the First Amendment.[367] As the now-ubiquitous saying goes, "[F]ree speech is not the same as free reach."[368] These and other intermediate methods will respect speech while acknowledging other interests.

Finally, the substantive difference between proportionality-based and categorical approaches is likely ultimately overstated. The two approaches tend to converge over time:

> Just as rule-based approaches often see the edges of the rules rounded off when difficult cases are presented, so too do more open-ended and discretionary approaches (which is what the "proportionality" inquiry amounts to) evolve, for reasons of limits on the human or judicial capacity to deal simultaneously with too many unorganized options, into approaches more reliant on rules.[369]

The project of constructing online systems of free expression has only just begun. Facebook, for example, only made its Community Standards public in 2018.[370] To say that content moderation remains incoherent, seemingly ad hoc, and unpredictable is to say nothing more than free speech is messy and that constructing a system of free expression based on new information technology will take time.

Proportionality is well-suited to this process. No single theory of free expression can resolve the entire range of complex free expression

---

365. See supra section I.B.1.d.

366. See Whitney v. California, 274 U.S. 357, 377 (1927) (Brandeis, J., concurring).

367. See Robert C. Post, Subsidized Speech, 106 Yale L.J. 151, 157 (1996).

368. Renee DiResta, Free Speech Is Not the Same as Free Reach, WIRED (Aug. 30, 2018), https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach [https://perma.cc/PW64-3W2J].

369. Schauer, Exceptional First Amendment, supra note 71, at 55; see also Frederick Schauer, The Convergence of Rules and Standards, 2003 N.Z. L. Rev. 303, 319 (arguing that rules and standards will approach "the same point on the rules-standards continuum regardless of whether they were given rules or standards as their starting raw material").

370. Casey Newton, Facebook Makes Its Community Guidelines Public and Introduces an Appeals Process, Verge (Apr. 24, 2018), https://www.theverge.com/2018/4/24/17270910/facebook-community-guidelines-appeals-process (on file with the *Columbia Law Review*).

problems.[371] Instead of continuing the search for a unified theory of free speech, adapted to the platform era, the development of content moderation principles is better served by a series of "incompletely theorized agreements," where decisionmakers reach agreements on outcomes based on relatively narrow or low-level explanations.[372] In his seminal articulation of this concept, Professor Cass Sunstein argued it was best suited to contexts where decisions need to be made rapidly, in the face of intractable social disagreements on first principles,[373] and by decisionmakers "lack[ing] . . . democratic pedigree."[374] Perhaps no adjudicatory environment better encapsulates these qualities than private platforms writing rules applied millions of times a day defining what speech is allowed in the fast-changing internet ecosystem. Rather than resorting to highfalutin justifications based on free speech platitudes, which have proved unsatisfying,[375] the context-specific reasoning style of proportionality serves to transform "a debate over values into a debate over facts, which is easier to resolve."[376]

This also better facilitates translation of speech rights into their online manifestation. Because rights contain "a significant empirical component, our understanding of a right can always be upset by evidence that forces a change in these empirical beliefs."[377] There can hardly be a more radical upset of our understanding of speech than the internet, a transformative force that is not static but continually evolving. Therefore, content moderation rules should be devised experimentally, "rather than implement[ed] . . . based on intuitive appeal," especially as empirical research has shown counterintuitive results.[378] Indeed, it is striking how much we do not know about online speech. Labels and fact-checking can be effective in certain circumstances, for example,[379] but their exact design

371. See Stone, Comparative Freedom of Expression, supra note 72, at 416 ("[T]he sheer complexity of the problems posed by a guarantee of freedom of expression makes it unlikely that a single 'theory' or 'set of values' might be appropriate for the entire range of freedom of expression problems.").

372. Cass R. Sunstein, Incompletely Theorized Agreements, 108 Harv. L. Rev. 1733, 1736 (1995).

373. Id. at 1735.

374. Id. at 1753.

375. See Mark Zuckerberg Stands for Voice and Free Expression, supra note 70.

376. Moshe Cohen-Eliya & Iddo Porat, Proportionality and the Culture of Justification, 59 Am. J. Compar. L. 463, 471 (2011).

377. T.M. Scanlon, The Difficulty of Tolerance: Essays in Political Philosophy 154 (2003).

378. Nicholas Dias, Gordon Pennycook & David G. Rand, Emphasizing Publishers Does Not Effectively Reduce Susceptibility to Misinformation on Social Media, 1 Harv. Kennedy Sch. Misinfo. Rev., Jan. 2020, at 1, 3.

379. See Ethan Porter & Thomas J. Wood, Why Is Facebook So Afraid of Checking Facts?, WIRED (May 14, 2020), https://www.wired.com/story/why-is-facebook-so-afraid-of-checking-facts [https://perma.cc/P68J-CQ74], and the studies cited therein.

really matters.[380] Without careful implementation, there can be unintended consequences like giving an "implied truth" effect to content left unlabeled (whether true or not).[381] We are only at the very beginning of the process of determining what works outside the take-down/leave-up paradigm.

One particularly important empirical question is the extent to which removing certain content can in fact be speech-enhancing, as people who might be chilled by the presence of harassment or abuse feel more empowered online.[382] Further research on this dynamic could transform the way we think about "protecting speech" online.

These few examples are simply to illustrate that content moderation needs to be grounded in research and not just reflexive interventions. Platitudes about the "marketplace of ideas" or "chilling effects" are in fact empirical assumptions based on very different speech ecosystems, unmediated by algorithms or shaped by the particular affordances of platforms.[383] For the first time in history, datasets exist to test these assumptions, and content moderation rules should be influenced and justified by reference to such findings.

This is no doubt a somewhat unsatisfying conclusion. Many particulars are left open for further research, which at the very least describes an uncertain future for speech rules, at least temporarily. This may confirm the fears of those who reject balancing precisely because it leads to more unpredictability than categorical, bright-line rules. But legal certainty is an instrumental value and should not be pursued at all costs. The value of strict rules versus flexible principles depends on context: "[B]oth have advantages and disadvantages. Rules lead to less than optimal decisions, whereas principles provide less legal certainty."[384] So, "[e]ven if 'categorical' rules would result in fewer errors . . . [proportionality] may result in fewer 'serious' errors, or departures from a common sense of

---

380. See Emily Saltz, Tommy Shane, Victoria Kwan, Claire Leibowicz & Claire Wardle, It Matters How Platforms Label Manipulated Media. Here Are 12 Principles Designers Should Follow., Medium: The Startup (June 9, 2020), https://medium.com/swlh/it-matters-how-platforms-label-manipulated-media-here-are-12-principles-designers-should-follow-438b76546078 [https://perma.cc/3VJZ-A8MD].

381. See Gordon Pennycook, Adam Bear, Evan T. Collins & David G. Rand, The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings, 66 Mgmt. Sci. 4944, 4945–46 (2020).

382. See, e.g., Danielle Keats Citron & Benjamin Wittes, The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity, 86 Fordham L. Rev. 401, 420 (2017) ("Individuals have difficulty expressing themselves in the face of online assaults."); Citron, Restricting Speech, supra note 122, at 122 ("Cyber harassment is now widely understood as profoundly damaging to victims' expressive . . . interests."); Danielle Keats Citron & Jonathon W. Penney, When Law Frees Us to Speak, 87 Fordham L. Rev. 2317, 2319 (2019) ("[O]nline abuse has a profound 'chilling effect.'").

383. See Fallon, Nature of Constitutional Rights, supra note 304, at 88 ("[C]urrent [free speech] doctrine quite plainly rests on empirical, predictive assessments.").

384. Petersen, supra note 170, at 55.

constitutional justice, than its 'categorical' counterpart."[385] The certainty of the more absolutist early post-as-trumps frame for content moderation extracted a high price. Indeed, the very point of disruption is to make *uncertain* the previously certain but unsatisfactory paradigm. And in truth, certainty and predictability are likely to be illusory in the context of the rapidly evolving, ever-changing environment of online platforms.

B. *Addressing Systemic Balancing's Current Deficits*

Accepting that content moderation is a task of systemic balancing says little about how that approach should be applied in practice. Indeed, the premise of the incommensurability critique is that there is little constraint on decisionmakers who balance. Add in the concession that some proportion of cases will inevitably be wrong, and creating any guardrails for decisionmakers might seem hopeless.

But "[t]he weight to give the indeterminacy critique depends to an important degree on what proportionality review would replace."[386] The rights-as-trumps approach also did not provide any meaningful constraint for writing speech rules. The history of online speech governance is a history of platforms exercising essentially unconstrained discretion, creating ad hoc rules in response to particular crises.[387] Platforms are still largely just "making rules up."[388] My defense of balancing is ultimately limited and pragmatic: I do not defend the rationality of balancing, nor do I take a position on whether a structured form of decisionmaking can completely constrain decisionmakers. Unlike others, I do not wish to propose a mathematical "internet interest-balancing formula" or tie myself to quantifiable cost-benefit analysis.[389] Instead, I readily concede that content moderation, at least for the foreseeable future, will be messy and often wrong. But the best path forward lies in accepting and being open about this so that mistakes can be discovered and corrected, rather than merely ignored or assumed away. Experimentation is essential. The goal of acknowledging the systemic balancing inherent in content moderation is therefore not to provide a completely constraining adjudicative method, but to expose to public scrutiny the decisionmaking process already taking

---

385. Jackson, supra note 15, at 3155.

386. Id. at 3153.

387. See Gillespie, Custodians of the Internet, supra note 59, at 67 ("Revisions of the guidelines often come only in response to outcries and public controversies.").

388. Kettemann & Schulz, supra note 29, at 28.

389. Cf. Robert Alexy, Mart Susi's Internet Balancing Formula, 25 Eur. L.J. 213, 213–14 (2019) (comparing the Internet Balancing Formula proposed by Mart Susi and the Weight Formula developed by Robert Alexy); Cass R. Sunstein, Does the Clear and Present Danger Test Survive Cost-Benefit Analysis?, 104 Cornell L. Rev. 1775, 1780–84 (2019) (laying out a framework for analyzing the costs and benefits of speech); Mart Susi, The Internet Balancing Formula, 25 Eur. L.J. 198, 204–07 (2019) (proposing a mathematical formula for balancing fundamental rights online).

place so it can be subject to public argumentation, contestation, and disruption.[390]

The goal is not to completely constrain balancing but to make it more accountable and, thereby, legitimate.[391] Because substantive outcomes will always be fundamentally contestable and contested, the task is not to arrive at "right" answers. Instead, "In conditions of relatively widespread reasonable disagreement . . . legitimacy connote[s] respect-worthiness."[392] To issue respect-worthy decisions, users and society more broadly have general intuitions that some sort of "due process" should apply.[393] And this has been reflected in demands for such process from civil society and academics.[394] It is because of these demands and general due process intuitions that, despite having no legal obligation to provide procedural protections in content moderation, increasingly all platforms do: They all engage in varying degrees of transparency reporting, and have appeal systems of more or less robustness and effectiveness.[395] The almost immutable lifecycle of user-generated content platforms is that after a certain scale such institutions of transparency and appeals become unavoidable.[396] This shows the powerful hold that due process intuitions have on the public imagination, and also that platforms, in responding to them, recognize them as legitimate.

Although only high level, there are general principles that can make this process of contestation more productive. The following sections look at this question first from the perspective of those designing the frontline institutions—platforms—and then from the perspective of their regulators.

---

390. On proportionality as a method of just such a constraint, see Mattias Kumm, The Idea of Socratic Contestation and the Right to Justification: The Point of Rights-Based Proportionality Review, 4 Law & Ethics Hum. Rts. 141, 143 (2010).

391. Kaminski, supra note 209, at 1567 (describing accountability as the "central problem of collaborative governance" in the context of algorithmic decisionmaking and targeted at "producing legitimacy—that is, the public acceptability of a system").

392. Richard H. Fallon, Jr., Law and Legitimacy in the Supreme Court 8 (2018); see also Jody Freeman, The Private Role in Public Governance, 75 N.Y.U. L. Rev. 543, 666 (2000) ("At bottom, legitimacy is a synonym for public acceptability, regardless of how it might be measured.").

393. Cass R. Sunstein & Adrian Vermeule, The Morality of Administrative Law, 131 Harv. L. Rev. 1924, 1964 (2018) (describing how such intuitions in judges are responsible for how due process requirements have developed in administrative law).

394. See, e.g., Zittrain, Three Eras, supra note 21, at 9; Santa Clara Principles on Transparency and Accountability in Content Moderation, Santa Clara Principles, https://santaclaraprinciples.org [https://perma.cc/FL97-RXVC] (last visited Oct. 24, 2020).

395. See supra notes 211–214 and accompanying text.

396. See, e.g., Sarah Perez, TikTok to Open a 'Transparency Center' Where Outside Experts Can Examine Its Moderation Practices, TechCrunch (Mar. 11, 2020), https://techcrunch.com/2020/03/11/tiktok-to-open-a-transparency-center-where-outside-experts-can-examine-its-moderation-practices/?guccounter=1 [https://perma.cc/7AR8-VBBG].

1. *Systemic Balancing for Platforms.* — What does systemic balancing require of platforms? The central point is that this form of governance demands transparency and candor so that trade-offs can be meaningfully debated, experimented with, and ultimately accepted, or at least acquiesced to, including by those who disagree with substantive outcomes. Of course, "[c]alls for greater transparency in the critique of social media are so common as to be nearly vacant,"[397] and perhaps this Article is merely adding one more cri de coeur to the pile. But understanding the nature of the balancing tasks that platforms are engaging in allows the *kind* of transparency required to be articulated with more specificity.

First, rules need *justification*. Proportionality is intrinsically tied to a "culture of justification" that recognizes rules restricting rights as legitimate only if the rulemaker provides adequate reasons, and not merely by pointing to naked authority.[398] And within this framework, "The coin of the realm . . . is the scrutiny of justifications."[399] This is why "categorization and . . . balancing . . . involve two very different intellectual styles,"[400] and demand "different types of reasons."[401] Moving from taxonomist to grocer makes the legitimacy of rules contingent on the rulemaker providing reasons that articulate the purpose of rules, the reason why they pursue legitimate aims, and what interests have been recognized and evaluated.[402] Platforms can no longer simply point to legal authority to make whatever rules they want. As platforms decide which rights and interests they will dignify, they need to explain how they have done so in order for their justifications to be assessed. Greater transparency can enhance trust and legitimacy by demonstrating "awareness of the intellectual, logistic, and moral depth" of content moderation decisions.[403] No platform currently meets this need. Instead, as Bowers and Zittrain observe, "The inwards-

397. Gillespie, Custodians of the Internet, supra note 59, at 198.

398. Cohen-Eliya & Porat, supra note 376, at 466–67.

399. Stone Sweet & Mathews, Proportionality Balancing and Constitutional Governance, supra note 14, at 56–57.

400. Daniel A. Farber, The Categorical Approach to Protecting Speech in American Constitutional Law, 84 Ind. L.J. 917, 919 (2009); Sullivan, supra note 17, at 293.

401. Kumm, Constitutional Rights as Principles, supra note 123, at 578.

402. See, e.g., Barak, Constitutional Rights and Their Limitations, supra note 73, at 486 ("[B]alancing brings—rather than confusion—a sense of order and method into constitutional law analysis. It forces the judge to identify the relevant principles and to provide a justification for the right's limitation . . . . It forces the judge to expose, both to themselves and to others, their train of thought."); Douek, All Out of Proportion, supra note 168, at 560–61 ("[I]t is a strength of proportionality that it respects judicial authority to determine the meaning and scope of constitutional rights or limitations, while also injecting the inherently legislative role into the process of constitutional adjudication through the assessment of public policy objectives."); Petersen, supra note 170, at 57 ("If we accept this line of reasoning, then balancing has one significant advantage: It is more transparent than categorical argumentation. Balancing forces the judges to mention the factors that determine their decision explicitly. By contrast, categorical reasoning often entails implicit balancing." (footnotes omitted)).

403. Facebook Data Transparency Advisory Group, supra note 40, at 41.

looking, largely public relations-oriented content governance models so widely deployed today are unsatisfying."[404] Informed contestation around rules and whether they are based on relevant considerations can only take place once there is transparency.

Again, the natural appeal of these characteristics is evident from platforms' gradual voluntary iteration toward them in the face of mounting public and regulatory pressure, even without legal mandates to do so. Take, for example, the controversy around Twitter and Facebook's different treatment of President Trump's post that included the phrase "when the looting starts, the shooting starts."[405] Twitter placed the tweet behind a warning screen, while Facebook did not take any action.[406] Both platforms acknowledged after the fact that they needed to be better at explaining how and when they applied their rules. Twitter said that "[w]e will continue to be transparent in how we make our decisions and be open with our rationale . . . . If we can't explain and be confident in our determination, we will not label a Tweet."[407] Zuckerberg similarly agreed that "we should have a more transparent process about how we weigh the different values and equities at stake, including safety and privacy."[408] In both these cases then, neither platform reneged on its substantive decision, but both emphasized the importance of a transparent *process* and reasoning.

Second, such justificatory reasoning must not be mere ad hoc balancing, but should be *structured*. Recent years have seen greater transparency of rules and public contestation around their merits. But the absence of any forum, method, or way of compelling platforms to participate has meant that such discourse is rarely focused and driven largely by what garners sufficient public outrage. This is clearly an unsatisfactory basis for deciding rules about speech, which should at least in some measure be protected against mere popular opinion. Developing methodological standards can be helpful here: Funneling such disputes through "the trappings of legal norms, methods, and determinacy or quasi-determinacy" can be both a protection against mere public pressure but also a way of maintaining sociological legitimacy for pervasive interest balancing, especially when countermajoritarian.[409]

---

404. Bowers & Zittrain, supra note 19, at 5.

405. Cristiano Lima, Zuckerberg: Facebook Leaving Up Trump's "Shooting" Post, Politico (May 29, 2020), https://www.politico.com/news/2020/05/29/zuckerberg-facebook-leaving-up-trumps-shooting-post-290292 [https://perma.cc/3Q8D-RQ8E].

406. Id.

407. Twitter Safety, supra note 1.

408. Mark Zuckerberg, Facebook (June 5, 2020), https://www.facebook.com/4/posts/10111985969467901/?d=n (on file with the *Columbia Law Review*).

409. Fallon, Nature of Constitutional Rights, supra note 304, at 34.

This maps the evolution of proportionality in global constitutionalism. Proportionality is not just a fancy label for open-ended balancing.[410] Instead, it is a series of prescribed decisionmaking steps.[411] Its structured nature gives it a greater disciplining and rationalizing effect[412] and provides the framework through which the value judgments inherent in the process of balancing are made visible.[413] As Schauer argues in one of the most important defenses of the rationality of balancing: "[T]his structure of burdens and presumptions . . . explains why it is a mistake to treat a proportionality enquiry as equivalent to an open-ended decision on the balance of all reasons and all facts . . . [and] gives a proportionality enquiry a degree of constraint . . . ."[414] Discussed above are some of the challenges of simply transposing the structure of proportionality testing from a state-based adjudicatory environment to content moderation.[415] But there are still some central characteristics that can and should be adopted. Restrictions on speech need to be plainly and publicly stated in advance (legality); the purpose of these restrictions needs to be clearly articulated; and all the relevant interests that have been balanced need to be identified (legitimacy). Additionally, the restriction needs to be shown to actually further this purpose and the identified interests (suitability), and be no more extensive than necessary for doing so and in proportion with the importance of the aim (necessity and proportionality).[416] Stepping through each of these elements can help sharpen and focus the exact nature of disagreements.

Because these requirements are so inherent to the method, the concern is not that without transparency and structure, platforms *might* not be applying the proportionality framework correctly. Without transparency and structure, platforms *cannot* be applying the proportionality framework correctly.

There are still many open questions to be answered, including the level of generality for justifications. If proportionality cannot be applied in

---

410. See Schauer, The Constraining Role, supra note 359, at 308–09 ("There is a difference between the structured enquiry of proportionality review and an open-ended mandate simply to 'do the right thing', or 'take everything into account', or make the best decision on the 'balance of reasons.'" (footnote omitted)).

411. See Stone Sweet & Mathews, Proportionality Balancing and Constitutional Governance, supra note 14, at 3–5; Jackson, supra note 15, at 3098–99.

412. Dieter Grimm, Proportionality in Canadian and German Constitutional Jurisprudence, 57 U. Toronto L.J. 383, 397 (2007).

413. See id.

414. Schauer, The Constraining Role, supra note 359, at 309.

415. See supra section II.A.

416. Grégoire Webber, Proportionality and Limitations on Freedom of Speech, *in* The Oxford Handbook of Freedom of Speech (Adrienne Stone & Frederick Schauer eds.) (forthcoming 2021) (manuscript at 7), https://ssrn.com/abstract=3358273 (on file with the *Columbia Law Review*) [hereinafter Webber, Proportionality and Limitations] (describing the leading formulation of proportionality, but noting that there are differences across jurisdictions and in scholarly interpretations).

every case—as it surely cannot, given the issue of scale—should it be assessed at the level of each category of rules (for example, hate speech) or each subcategory (for example, white nationalism) or, for that matter, each region, country, or community? This is a hard question that has not been satisfactorily resolved in legal systems with experience applying these tests. Courts and decisionmakers are remarkably inconsistent about the requisite level of generality in areas as diverse as disparate impact law,[417] to proportionality analysis in general freedom of expression jurisprudence.[418] The best answer is likely to be that it depends on context: Questions about what constitutes "coordinated inauthentic behavior," which is ostensibly content-neutral,[419] for example, might be answered at a higher level of generality than what constitutes "hate speech," which is inherently context- and culture-specific.

Again, all these unknowns may suggest that fixating on reasoning structure as a constraint is hopelessly naïve. It seems to ignore everything legal realism has taught. But remember the currently completely unconstrained baseline. Structured reasoning does not and cannot promise to eliminate all subjectivity or manipulation toward preferred outcomes. It can, however, promise to make such manipulation more visible than the currently completely opaque and ad hoc systems. The goal is, as Justice Breyer has argued, to make "the calculus behind an opinion explicit so that it can be seen and criticized."[420] A failure to fully realize rule of law ideals should not lead to abandoning them entirely; there will be tradeoffs between "discretion, transparency, retroactivity, and intelligibility," and finding the optimal level will include considering costs in multiple directions.[421] The current reckoning around content moderation has been prompted by an emerging consensus that the level of platform discretion has been too high and the realization of the other important considerations too low. Balancing these considerations and the fact that a level of discretion will always persist is no reason to forsake the project.

Third, *error rates* need to be transparently acknowledged and defended. The unique challenges of online speech ecosystems force a probabilistic assessment of speech rights, but embarrassment about openly defending such decisions has forced consequential decisions about error-choice into the shadows. Platforms need to display the same candor about their anticipation of mistakes as consistently as they did at the start of the

---

417. Louis Kaplow, Balancing Versus Structured Decision Procedures: Antitrust, Title VII Disparate Impact, and Constitutional Law Strict Scrutiny, 167 U. Pa. L. Rev. 1375, 1430 (2019).

418. Webber, Proportionality and Limitations, supra note 416, at 14.

419. For discussion, see Evelyn Douek, The Free Speech Blind Spot: Foreign Election Interference on Social Media, *in* Defending Democracy: Combating Foreign Election Interference in a Digital Age 265, 269–72, 277 (Duncan B. Hollis & Jens David Ohlin eds., 2021).

420. Breyer, supra note 305, at 257.

421. Sunstein & Vermeule, supra note 393, at 1974.

COVID-19 pandemic.[422] The pandemic example proves the potential value of transparency and candor as well. As a result of being upfront about the unavoidable reason of increased reliance on AI tools and the resultant expectation of more mistakes, even the civil libertarian thinktank Electronic Frontier Foundation—normally highly critical of AI moderation—gave platforms credit for their frankness, accepted the decision in context, and focused on exhorting platforms to ensure such measures were temporary.[423] No doubt, many have been critical of platforms' content moderation during the pandemic, but such criticism is not due to transparency. There will always be criticism of content moderation; it pre-dates platforms' increased transparency in recent years and will persist even if platforms meet all demands for greater disclosure. But it is only with transparency that such criticisms can be informed ones.

Transparency also has costs and should not be unduly reified.[424] It consumes resources, and it can reduce candor or increase scrutiny, thereby reducing trust. The answer to this is that while there may be trade-offs, the baseline of near total opacity is unlikely to be the right answer. As has been noted in another context, "Reasonable minds may differ as to what the ideal balance . . . might be, but it seems unlikely that this balance should be left to the judgment of a private corporation."[425]

A corollary of error acceptance is the need for a way to challenge and rectify mistakes. Mistakes are inevitable, but not always acceptable. No doubt part of the reason platforms do not openly acknowledge their error choices is because they have failed to build adequately robust systems for error correction.[426] The failure to provide adequate procedural checks is not separate but related to the dissatisfaction with the substantive rules. Mistakenly removing volunteer mask-makers or antiracist skinheads might be more readily acceptable, for example,[427] if a reliable process existed for ensuring such mistakes were indeed temporary rather than relying on media outrage to force reversals.

Finally, proportionality requires that enforcement be the *least restrictive means*.[428] This is inextricable from proportionality and cannot be

422. See supra section I.D.2.

423. Jillian C. York & Corynne McSherry, Automated Moderation Must Be Temporary, Transparent and Easily Appealable, Elec. Frontier Found. (Apr. 2, 2020), https://www.eff.org/deeplinks/2020/04/automated-moderation-must-be-temporary-transparent-and-easily-appealable [https://perma.cc/3H4F-GGX7].

424. See David E. Pozen, Seeing Transparency More Clearly, 80 Pub. Admin. Rev. 326, 326 (2020) ("Transparency, that is, has been identified as the cause of, and solution to a remarkable range of problems.").

425. Comment, Cooperation or Resistance?: The Role of Tech Companies in Government Surveillance, 131 Harv. L. Rev. 1722, 1729 (2018).

426. For one influential example of longstanding calls for greater due process and opportunities for appeals, see Santa Clara Principles on Transparency and Accountability in Content Moderation, supra note 394.

427. See supra section II.B.

428. See supra section I.B.1.d.

a mere afterthought. The take-down/leave-up binary of content moderation must be consigned to the posts-as-trumps era, and more imagination about other interventions is needed, as described above.[429] And as platforms experiment more with interventions like labels, warning interstitials, nudges to read before sharing, or the ability to block or hide certain users or replies, they need to be transparent about the results of these measures so that their true (and not merely intuitive) proportionality and probabilities can be assessed.

An example of successful and accepted systemic balancing might be helpful: spam. Spam filters, which rely on block lists and Bayesian filters, are inherently probabilistic.[430] As any email user can attest, spam filters regularly make mistakes. Spam is consistently one of the largest categories of content removal for every platform.[431] Early on, there were concerns about the free speech implications of the probabilistic approach of spam filters given the propensity to block and chill some legitimate speech.[432] The controversy was never really resolved, but spam filtering is now generally accepted as an "essential part of the internet."[433] Indeed, "Removing spam is censoring content; it just happens to be content that nearly all users agree should go."[434] Here, the costs of spam speech, and especially the potential to flood non-spam content, have been accepted as outweighing the individual right to spam and the inevitable errors of spam filters. This has no doubt been aided by increasingly accurate spam filters—that is, better probabilities. Therefore, spam filters impose a probabilistic tax on internet free speech, but they have gained acceptance as a proportionate one.[435]

2. *Systemic Balancing for Regulators.* — Governments themselves cannot perform the balancing or enforcement for all content moderation rules in the form of regulation. In many cases, content moderation extends far beyond the limits of a government's constitutional power to regulate speech. But even in those cases where the government could, should, and increasingly will, proscribe certain speech, the scale of enforcement, the technical requirements for doing so, and the diversity and pace of change

---

429. See supra section I.B.2.c.

430. Finn Brunton, Spam: A Shadow History of the Internet 133–55 (2013).

431. See supra notes 211, 213–214.

432. See, e.g., Cindy Cohn & Annalee Newitz, Noncommercial Email Lists: Collateral Damage in the Fight Against Spam, Elec. Frontier Found. (Nov. 12, 2004), https://www.eff.org/wp/noncommercial-email-lists-collateral-damage-fight-against-spam [https://perma.cc/6XB8-BUU8] ("Although ISPs may have the best of intentions, what we see in this scenario—one that is all too common—is free speech being chilled in the service of blocking spam.").

433. Sarah Jeong, The Internet of Garbage 67 (2018).

434. Gillespie, Custodians of the Internet, supra note 59, at 217 n.22.

435. Renee DiResta, The Return of Fake News—and Lessons from Spam, WIRED (June 5, 2019), https://www.wired.com/story/the-return-of-fake-news [https://perma.cc/5MN2-92LW] ("Today, the vast majority of email that's clearly crap is stopped at the source—and no one mourns the free speech rights of spammers.").

of the online speech environment will always leave platforms as the primary regulators of online speech. The role of government regulators is therefore to institutionalize, incentivize, and verify the systemic balancing of platforms.[436] Recognizing this illuminates some persistent points of contention with proposals for regulatory reform in this area.

The first is that a punitive approach to regulation that punishes content moderation systems for individual errors is unrealistic and creates bad incentive problems. Because perfect enforcement is impossible, such an approach irreversibly puts a finger on the content moderation balancing calculus by making the platforms' own interest significantly more weighty. This is just a recasting of the long-noted issue that online intermediaries are the "weakest link" in protecting speech because they have limited incentive to avoid over-censorship.[437] The marginal benefit platforms receive from allowing any particular post to stay up is minimal, so potential liability for failure to remove even a single piece of offending content makes balancing simple. Experience with such laws "tells us that when platforms face legal risk for user speech, they routinely err on the side of caution and take it down."[438] Putting this in terms of systemic balancing: When regulators try to enshrine a particular balancing outcome into law, they can in fact change that balancing by tipping the scales of those that enforce it (the platforms) toward a higher probability of over-enforcement.

The problem cuts both ways. A focus on individual errors also risks misdiagnosing the true problems with content moderation system design.[439] As a report commissioned by the French Government noted, in the absence of sufficient data about how content moderation systems actually work, "the public authorities and representatives of civil society are reduced to highlighting individual examples of unmoderated or poorly moderated content. Yet these isolated failures are insufficient to prove a potential systemic failure."[440] And as Professors Nathaniel Persily and Joshua Tucker note, "[I]t is exceedingly easy (and often misleading) to find cursory evidence of *anything* on social media because there is so much

436. In this way, their role is akin to systems of collaborative governance described in Kaminski, supra note 209, at 1562.

437. Seth F. Kreimer, Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link, 155 U. Pa. L. Rev. 11, 27 (2006) ("The strategy of recruiting proxy censors by targeting the weakest link in the chain of communication has obvious advantages for regulators.").

438. Keller, Observations, supra note 225, at 2.

439. See Kaminski, supra note 209, at 1580 ("[W]here systemic and individual accountability can be complementary, there is also a danger of confusing one kind of accountability for another and crafting a system that is accountable along only one axis."); Andrew D. Selbst & Solon Barocas, The Intuitive Appeal of Explainable Machines, 87 Fordham L. Rev. 1085, 1130–33 (2018) (describing how evaluating decisionmaking that relies on a machine learning model requires scrutiny of the institutional and subjective process behind its development).

440. French Sec'y of State for Digit. Affs., supra note 35, at 12.

of it . . . . The more important questions . . . require much more complex research . . . ."[441] Of course, platforms themselves are to blame for the current lack of systemic transparency, and for failing to sufficiently make the case for error rates being an inherent part of content moderation.[442] But as long as transparency remains voluntary, there is an incentive problem: The more information an individual platform makes available, the more scrutiny can be applied and the more criticism can be made. Transparency mandates enshrined in law are therefore crucial to make systemic accountability consistent across platforms. This is no doubt part of why bigger platforms, like Facebook, which have instituted transparency measures as a result of public pressure, advocate for them becoming enshrined in law.[443] An important caveat is that the extent of mandated transparency should be made relative to a platform's size, so that regulatory compliance does not become a barrier to entry.

The second lesson for regulators is that there will always be categories of content moderation rules that private platforms can and should make that will be beyond the reach of the law to substantively regulate. This does not mean that there are *no* relevant speech interests; the balancing calculus is simply different. When the state is involved, the distrust of speech regulation is at its highest,[444] the tools are at their bluntest, and the decisionmaking processes are at their slowest and most rigid. State actors will always have a role in defining illegal content, subject to the constraints of their particular constitutional system. Beyond this, state regulation of "legal but harmful" content is inherently fraught.[445]

Therefore, third, the question is how to expand and legitimate these private systems of content moderation where government cannot reach but platforms should still be accountable.[446] But how can the balancing of incommensurables premised on the inevitability of errors be accountable?

Here, some intellectual modesty is required. Content moderation is in its adolescence. In its youth, it was dominated by a categorical, posts-as-

---

441. Nathaniel Persily & Joshua A. Tucker, Conclusion: The Challenges and Opportunities for Social Media Research, *in* Social Media and Democracy: The State of the Field, Prospects for Reform 313, 324 (Nathaniel Persily ed., 2020).

442. A notable exception, outside the context of the pandemic, is Bickert, supra note 32, at 7 ("Enforcement will always be imperfect.").

443. Id. at 11 ("An important benefit of this regulatory approach is that it incentivizes an appropriate balancing of competing interests, such as freedom of expression, safety, and privacy. Required transparency measures ensure that the companies' balancing efforts are laid bare for the public and governments to see.").

444. Ely, supra note 78, at 106 ("Courts must police inhibitions on expression and other political activity because we cannot trust elected officials to do so: ins have a way of wanting to make sure the outs stay out.").

445. Graham Smith, Online Harms IFAQ*, Cyberleagle (Feb. 2, 2020), https://www.cyberleagle.com/2020/02/online-harms-ifaq.html [https://perma.cc/M8WG-CBXA].

446. As recommended by the French Secretary of State for Digital Affairs, supra note 35, at 11.

trumps approach. It is now in a period of experimentation on its way to a more mature paradigm. But, as has been repeatedly emphasized, these laboratories of online governance cannot be evaluated or legitimated without transparency.[447] Therefore the role of public regulation can be to turn the inward-looking and unsatisfying systems of content regulation outward[448] and focus on making them more accountable and credible.[449] Simple disclosures from platforms will not be enough. Oversight and validation of the systems and figures that platforms invoke to demonstrate their responsible behavior is also necessary. In this period of experimentation, regulators should focus on verifying that content moderation systems do what they say on the tin. When platforms report ever greater takedowns of content and proactive detection by AI tools, regulators should independently audit these figures to verify these numbers and ensure they are not simply a result of decreases in decision quality. Such oversight can also help identify blind spots in company processes, such as Facebook's civil rights audit, which identified simple changes in the interface content moderators use to increase accuracy.[450]

A model for this kind of approach identified by the French government report is the financial sector, where independent audits focus on verifying that banks have appropriate systems in place, rather than punishing discrete instances of unlawful use of their services.[451] The obvious merit of this model, aside from focusing on the systemic nature of content moderation, is that it does not require government involvement in substantive decisions about rules for online speech, but rather focuses on regulating systems and processes.

The final lesson is to avoid lock-in. Every single choice in this context involves difficult trade-offs that will have unknown dynamic interactions not only within individual platforms, or across the broader internet ecosystem, but also with changes and developments in norms, expectations, and behaviors in the offline world. Oversight models need to provide

---

447. See, e.g., Evelyn Douek, YouTube's Bad Week and the Limitations of Laboratories of Online Governance, Lawfare (June 11, 2019), https://www.lawfareblog.com/youtubes-bad-week-and-limitations-laboratories-online-governance [https://perma.cc/BM9H-2WLT].

448. See Bowers & Zittrain, supra note 19, at 5 ("[R]esponsibility for key aspects of content governance must take place at least in part outside of the platforms, at an organizational remove from their business interests.").

449. French Sec'y of State for Digit. Affs., supra note 35, at 10 (suggesting methods for improving accountability and transparency).

450. Facebook's Civil Rights Audit, supra note 25, at 43; Evelyn Douek, Facebook Releases Civil Rights Audit Progress Report, Lawfare (July 1, 2019), https://www.lawfareblog.com/facebook-releases-civil-rights-audit-progress-report [https://perma.cc/CGT7-3R3Y].

451. See Evelyn Douek, Two Calls for Tech Regulation: The French Government Report and the Christchurch Call, Lawfare (May 18, 2019), https://www.lawfareblog.com/two-calls-tech-regulation-french-government-report-and-christchurch-call [https://perma.cc/8QFX-AY4T] (summarizing the French Government report).

room to adapt and learn as these interactions become apparent and change. Focusing on transparency and procedural rights to challenge decisions and disrupt the status quo are the likeliest to enable this. These systems and platforms are still incredibly young—regulations that create path dependence and treat current platform configurations as inevitable only risk entrenching them.

C. *Returning from the Not-So-Exceptional State of Exception*

One question will dominate content moderation's return to "normal" from its pandemic state of exception: If platforms could and did police misinformation more aggressively during a public health emergency, why don't they do this all the time? Commentators began asking this question as soon as platforms' state of emergency was declared.[452] And, indeed, in certain cases the creep has already started.[453]

The question is intuitive: For a long time, platforms resisted calls to remove content on the basis of falsity alone.[454] Famously, they insisted they should not be "arbiters of truth,"[455] categorically refusing to take content down because it is untrue. Then the pandemic caused platforms to change their position, almost literally overnight. Now, Facebook is removing false claims or conspiracy theories that have been flagged by authorities and that could cause harm.[456] Twitter has a long list of types of false claims it will remove[457] and has started labeling tweets that falsely link 5G to COVID-19.[458] The list of misinformation that YouTube is removing is similarly broad, including categories that could arguably be political, such as content disputing the efficacy of WHO or local health authority guidelines

---

452. See Douek, The Internet's Titans, supra note 13 (questioning whether tech companies' expanded content regulation powers in light of the pandemic will become the new "normal"); see also Joan Donovan, You Purged Racists from Your Website? Great, Now Get to Work, WIRED (July 1, 2020), https://www.wired.com/story/you-purged-racists-from-your-website-great-now-get-to-work [https://perma.cc/99D9-FUMF] (arguing that companies should "take decisive action to control who and what is on their sites"); Alex Kantrowitz, Facebook Is Taking Down Posts that Cause Imminent Harm—But Not Posts that Cause Inevitable Harm, BuzzFeed News (May 23, 2020), https://www.buzzfeednews.com/article/alexkantrowitz/facebook-coronavirus-misinformation-takedowns [https://perma.cc/M9PS-PJRU] (noting that some would like Facebook to "consistently" and "aggressively" continue its crackdown).

453. Douek, Trump Is Banned, supra note 146.

454. See supra notes 85–90 and accompanying text.

455. See, e.g., Borchers, supra note 86; Srinivasan, supra note 86; Zuckerberg, Status Update, supra note 86.

456. Jin, supra note 277.

457. Gadde & Derella, supra note 279.

458. Jon Porter, Twitter Starts Aggressively Fact-Checking Tweets Linking 5G to COVID-19, Verge (June 9, 2020), https://www.theverge.com/2020/6/9/21284940/twitter-5g-coronavirus-covid-19-fact-check-disinformation-conspiracy-theories-label (on file with the *Columbia Law Review*).

on physical distancing measures.[459] These platforms have proven they *can* remove misinformation—so why shouldn't they do so across the board?

First, even if "can" *did* necessarily imply "ought," the "can" has not been demonstrated. There is no meaningful data on how effective platforms have been at enforcing their rules during the pandemic. There has been plenty of rhetoric about a firmer hand, but actual results remain opaque. The only thing that is certain is that policies are being both over- and under-enforced in certain cases. The extent of each, and in what ways, however, remains unknown. We currently know the proportionality of the rules that platforms formulated to deal with the pandemic, but we do not know any of the probabilities with which they have been enforced. There is no particular reason to assume that platforms have been unusually successful, given their poor track record on enforcement generally. But the pandemic created a forced experiment of heavy-handed, AI-dependent content moderation, and the comparative results could usefully inform the debate over content moderation's future. A first step for regulators should be forcing companies to share data so claims of efficacy can be evaluated. There are difficult privacy trade-offs here, and an exact regulatory model is beyond this Article's scope. But laws clarifying privacy carve-outs for research and imposing obligations on the largest "data stewards" in society to contribute to public-facing research would be a start.[460]

Even if one assumes that content moderation during the pandemic has largely been a success story, there remains the question of what to do once the public health emergency subsides. The invocation of emergency powers was welcome during the pandemic, but accepting such measures across all categories of content at all times would in fact significantly augment the unaccountable power that these companies exercise over public discourse.[461] This is not a situation that should merely be stumbled into without introducing proper safeguards.

It is beyond the scope of this Article to answer whether the treatment of health misinformation is a good model for the treatment of false speech more generally. This Article aims only to illustrate the terms on which this debate should take place, and show that the debate itself is one about proportionality and probability.

Categorical reasoning makes the question one of definition and line-drawing: Can you delineate between health misinformation and other

459. COVID-19 Medical Misinformation Policy, supra note 283.

460. Persily & Tucker, supra note 441, at 322–23.

461. See Douek, COVID-19 and Social Media Content Moderation, supra note 39 (arguing that "the public is asking tech platforms to step up [during the pandemic], but we also need to keep thinking about how to rein them in"); Douek, The Internet's Titans, supra note 13 ("What's happening during the pandemic is just an accentuated version of the norm. It has shown that even the most seemingly entrenched rules can be instantly overthrown. Right now, this may be helpful. But what about once the worst of this crisis subsides?").

kinds of misinformation? The pandemic has made clear that categorical boundaries between health and politics are blurry at best.[462] But treating all misinformation alike creates problems. Platforms had justified their more heavy-handed approach to content moderation during the pandemic, in part, based on the existence of more accepted "authoritative" sources of information, such as the WHO.[463] To what authorities can platforms appeal as having legitimacy to decide the truth or falsity of political claims?[464] This can be described in terms of error rates: In the context of medicine, the category of content is easier to detect, and the evidentiary standards necessary to establish falsity are much easier to meet (although, as we have seen, by no means always easy[465]). As political scientists Sarah Kreps and Brendan Nyhan observe, "Standards of truth and accuracy in politics are more subjective and likely to provoke controversy."[466] The categorical framework also does not provide guidance on how to think about the very different nature of the harms involved in health and political misinformation, and the ways they might require different responses. The most effective and least restrictive means may be different.

Proportionality and probability allow for a process of reasoning that can account for the different interests involved, their different weights, and the different error rates in enforcement across these different categories of content. This type of reasoning can also suggest responses that account for these differences and do not try to force the answer into the box of either categorically staying up or coming down. Finally, proportionality- and probability-based reasoning forces the question of whether the error rates that might seem acceptable in the context of a public health emergency are the same as those that are acceptable all the time.

---

462. See, e.g., Isaac Stanley-Becker, Mask or No Mask? Face Coverings Become Tool in Partisan Combat., Wash. Post (May 12, 2020), https://www.washingtonpost.com/politics/in-virus-response-riven-by-politics-masks-are-latest-rorschach-test/2020/05/12/698477d4-93e6-11ea-91d7-cf4423d47683_story.html (on file with the *Columbia Law Review*) (explaining how decisions about mask wearing and other coronavirus responses have been, in part, politically motivated).

463. Facebook Press Call, supra note 10, at 17 ("[T]here are broadly trusted authorities who people across . . . society would all agree can arbitrate . . . what's trustworthy and what's not . . . .").

464. On the difficulties of this, see, for example, Anton Troianovski, Fighting False News in Ukraine, Facebook Fact Checkers Tread a Blurry Line, N.Y. Times (July 26, 2020), https://www.nytimes.com/2020/07/26/world/europe/ukraine-facebook-fake-news.html (on file with the *Columbia Law Review*).

465. Renee DiResta, Health Experts Don't Understand How Information Moves, Atlantic (May 6, 2020), https://www.theatlantic.com/ideas/archive/2020/05/health-experts-dont-understand-how-information-moves/611218 (on file with the *Columbia Law Review*) ("Determining who is an authoritative figure worth amplifying is more challenging than ever.").

466. Sarah Kreps & Brendan Nyhan, Coronavirus Fake News Isn't Like Other Fake News, Foreign Affs. (Mar. 30, 2020), https://www.foreignaffairs.com/articles/2020-03-30/coronavirus-fake-news-isnt-other-fake-news (on file with the *Columbia Law Review*).

A categorical frame obscures all this complexity, but determining what content moderation should look like after the crisis is no simple task. A lens of systemic balancing will at least ask the right questions.

CONCLUSION

Content moderation during the pandemic has simply made more apparent what is an ever-present truth: Platforms' systems are governed by a logic of systemic balancing, where decisions are made on the basis of their proportionality and probabilities. Recognition of these precepts has been slow, not least because they seem at odds with the traditional categorical and individualistic framing of speech rights that dominates the First Amendment tradition from which the major tech platforms hail. But openly acknowledging these dynamics is the first step toward recalibrating around them in the internet's adolescence, where laboratories of online governance must experiment to find acceptable and accepted solutions to the intractable challenges of online speech governance.

Successful online speech governance is not an end point to be arrived at, but an ongoing project of iteration, calibration, and explanation based on changing rules, norms, and technical capacity. In this dynamic environment and new paradigm of private governance of publicly important speech, to subject online speech rights to proportionality analysis or systemic optimization is not to deny those rights altogether.[467] In the modern platform era, it is the only way to legitimately realize them at all.

467. Cf. Dworkin, supra note 51, at xi, 192 (1977). As Dworkin states, "Individuals have rights when, for some reason, a collective goal is not a sufficient justification for denying them what they wish, as individuals, to have or to do." For it to be otherwise would "make [a] claim of a right pointless." Id.