VOLUME 136 DECEMBER 2022 NUMBER 2

HARVARD LAW REVIEW

ARTICLES

CONTENT MODERATION AS SYSTEMS THINKING

Evelyn Douek

CONTENTS

INTRODUCTION 528
I. THE STANDARD PICTURE OF CONTENT MODERATION 535
II. THE STANDARD PICTURE'S BLIND SPOTS 539
A. Content Moderation Bureaucracies Are a "They" Not an "It" 539
1. Non-Content-Based Content Moderation 539
2. Cross-Platform and Government Cooperation 542
3. Delegated Decisionmaking 543
4. Design and Affordances 545
B. Content Moderation Is All About Tradeoffs 548
1. Error Choices: False Positives Versus False Negatives 548
2. Gatekeeping Versus User Control 552
3. Procedural Justice and Legitimacy Tradeoffs at Scale 553
4. Consistency Versus Contextual Decisionmaking 554
C. The Standard Picture's Mistaken Assumptions 556
1. Speech Is Not So Special 556
2. Ex Post Review Can Be Systemic 563
3. The Pursuit of Perfectibility Is Misguided 564
III. MISGUIDED REFORM EFFORTS BASED ON THE STANDARD PICTURE 564
A. The Standard Picture's Influence 565
B. The Futility and Failures of Individualized Ex Post Review 568
1. Individual Errors Are Poor Vehicles for Identifying Systemic Failures 569
2. Case-by-Case Review Provides Inadequate Remedies 571
3. Transparency Theater 572
4. Process Theater 577
IV. SECOND WAVE CONTENT MODERATION INSTITUTIONAL DESIGN 584
A. Structural Mandates 586
1. Separation of Functions 586
2. Complaints Mechanisms 589

3. Disclosure of Nature and Extent of Contacts with Third-Party Decisionmakers 591
4. Retention of and Provision for Access to Data 592
B. Procedural Requirements 593
1. Annual Content Moderation Plans and Compliance Reports 593
2. Quality Assurance and Auditing 601
3. Aggregated Claims 602
C. Enforcement 603
CONCLUSION: MOVING SLOWLY AND FIXING THINGS 606

CONTENT MODERATION AS SYSTEMS THINKING

Evelyn Douek[*]

The stylized picture of content moderation that forms the basis for most regulatory and academic discussion of online speech governance is misleading and incomplete. That picture depicts content moderation as a rough online analog of offline judicial adjudication of speech rights, with legislative-style substantive rules being applied over and over again to individual pieces of content by a hierarchical bureaucracy of moderators. This understanding leads regulators and scholars to assume that the best way to make platforms accountable for their decisions about online speech is to ensure platforms provide users the kind of ex post individual review provided by courts in First Amendment cases and to guarantee users with ever more due process rights. But because the scale and speed of online speech means content moderation cannot be understood as simply the aggregation of many (many!) individual adjudications, what this approach produces is accountability theater rather than actual accountability. This Article argues that content moderation should instead be understood as a project of mass speech administration and that looking past a post-by-post evaluation of platform decisionmaking reveals a complex and dynamic system that needs a more proactive and continuous form of governance than the vehicle of individual error correction allows. Lawmakers need to embrace a second wave of regulatory thinking about content moderation institutional design that eschews comforting but illusory First Amendment–style analogies and instead adopts a systems thinking approach. This approach focuses on the need to look to structural and procedural mechanisms that target the key ex ante and systemic decisionmaking that occurs upstream of any individual case.

> *I'll let you write the substance . . . and you let me write the procedure, and I'll screw you every time.*
>
> — Congressman John Dingell[1]

INTRODUCTION

The stylized picture of content moderation that forms the basis for most regulatory and academic discussion of online speech governance is misleading and incomplete.[2] This standard picture depicts

[*] Assistant Professor, Stanford Law School. Many (many!) thanks to Elena Chachko, Julie Cohen, Ros Dixon, Brenda Dvoskin, Jody Freeman, Tarleton Gillespie, Katy Glenn Bass, Jack Goldsmith, Robert Gorwa, James Grimmelmann, Woody Hartzog, Jameel Jaffer, Quinta Jurecic, Thomas Kadri, Daphne Keller, Ramya Krishnan, Genevieve Lakier, Dina Lamdany, Paddy Leerssen, Emma Llansó, Martha Minow, Anne Joseph O'Connell, Nate Persily, Andrew Selbst, Rory Van Loo, Ari Ezra Waldman, and Jonathan Zittrain. Thank you also to the wonderful editors of the *Harvard Law Review*, who were a dream to work with. Errors and blind spots, both individual and systemic, entirely my own. This one's for my mum.

[1] *Regulatory Reform Act: Hearing on H.R. 2327 Before the Subcomm. on Admin. Law and Governmental Reguls. of the House Comm. on the Judiciary*, 98th Cong. 312 (1983) (statement of Rep. John Dingell, Chairman, H. Comm. on Energy & Com.).

[2] I use "content moderation" to mean platforms' systems and rules that determine how they treat user-generated content on their services. This generally accords with Professor James Grimmelmann's definition. *See* James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 47 (2015) (defining "moderation" as "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse").

content moderation as a process in which social media platforms write a set of legislative-style substantive rules and apply them in individual cases. This picture leads regulators and scholars to assume that the most effective form of error correction and accountability for content moderation decisions is individual ex post review. If a post is mistakenly taken down for depicting breasts but actually depicted onions,[3] the error can be reversed by user appeal. If a post breaks a platform's rules for inciting violence but was left up, a user can flag the post for further review to be removed. This model is a familiar way of thinking about speech disputes: the standard picture of content moderation bears a striking resemblance to the legalistic way speech rights work offline in the First Amendment context. In this picture, platforms are "The New Governors," constructing governance systems similar to the offline justice system in which "[c]ontent moderators act in a capacity very similar to that of judges."[4]

This picture of content moderation has significant implications for how regulation of online speech is designed. Because it focuses on the merits of individual speech decisions, it leads to endless and irresolvable arguments about the normative desirability of platforms' substantive rules, whether they have been correctly and impartially applied in particular cases, and whether platforms have afforded due process to individual users. These are the questions that dominate media headlines: should Facebook[5] prohibit Holocaust denial and can it even enforce such a ban if it wanted to?;[6] was Twitter right to suspend President Donald Trump's account and did it give him adequate notice?;[7] should YouTube have rules prohibiting content delegitimizing election results, like other major platforms do?[8] Many legislative efforts worldwide to

[3] Yes, this is a real example: *Why Some Onions Were Too Sexy for Facebook*, BBC NEWS (Oct. 8, 2020), https://www.bbc.com/news/54467384 [https://perma.cc/TUM8-QYFT].

[4] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1647 (2018). Professor Kate Klonick was inspired by a workshop by Professor Rory Van Loo for a paper in which he had more broadly portrayed the corporation as acting like a judge and shaping "the de facto substantive rules governing the vast majority of consumer disputes." Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REGUL. 547, 554, 566, 602 (2016) ("Corporations are increasingly assuming roles associated with courthouses." *Id.* at 554.); *see* Klonick, *supra*, at 1599 n.*.

[5] Because this Article was written and predominantly refers to events that occurred before Facebook changed its name to Meta, for consistency it refers in text to both the platform and its parent company as Facebook.

[6] *See* Aaron Sankin, *Facebook Said It Would Ban Holocaust Deniers. Instead, Its Algorithm Provided a Network for Them*, THE MARKUP (Nov. 24, 2020, 8:00 AM), https://themarkup.org/news/2020/11/24/facebook-ban-holocaust-deniers-antisemitism [https://perma.cc/2CYY-KUGN].

[7] *See* Twitter, Inc., *Permanent Suspension of @realDonaldTrump*, TWITTER BLOG (Jan. 8, 2021), https://blog.twitter.com/en_us/topics/company/2020/suspension.html [https://perma.cc/2RKG-6F9L].

[8] *See* Casey Newton, *How YouTube Failed the 2020 Election Test*, PLATFORMER (Mar. 3, 2021), https://www.platformer.news/p/how-youtube-failed-the-2020-election [https://perma.cc/G6QN-QF5T].

rein in platform power over public discourse similarly focus on these substantive questions.

But in fact, many of the most important decisions in content moderation happen outside and upstream of this standard picture, before any individual content moderation case even arises. This Article's central claim is that the standard picture's focus on the treatment of individual posts is misguided and that the toolset for content moderation reform needs to be expanded beyond individual error correction. It advocates for a systems thinking approach to content moderation regulation that focuses on systems rather than individual cases, on wholes and interrelationships rather than parts, and on "patterns of change rather than static snapshots."[9] Such an approach acknowledges that individual errors may be the canary in the coal mine of systemic failure but are not by themselves evidence of inadequate content moderation systems and that rectifying such errors will not bring overall accountability to regulators or the public.[10] Looking past a post-by-post evaluation of platform decisionmaking reveals a complex and dynamic system that needs a more proactive and continuous form of governance than the vehicle of individual error correction allows.

Understanding this broader picture of content moderation is crucial now because regulation is on its way (and, in some cases, already here). There is growing consensus that the rules for what people can say online are too important to leave entirely to private actors who have no formal legal obligation to defend those rules' rationality or enforce those rules with any consistency, and who are otherwise unaccountable to lawmakers and the public. But designing laws to bring accountability to content moderation based on the standard picture would be a mistake for two primary reasons.

First, the standard approach fails to reckon with the ways in which the scale and speed of online speech governance are fundamentally different from offline speech governance.[11] "Content moderation" is not just the aggregation of many (many!) binary decisions to take down or

[9] Ross D. Arnold & Jon P. Wade, *A Definition of Systems Thinking: A Systems Approach*, 44 PROCEDIA COMPUT. SCI. 669, 671–74 (2015) (reviewing the many definitions of "systems thinking").

[10] This Article adopts a thin version of accountability, meaning simply checks on decisionmaking, in the form of requiring one party to provide information to another party, that are intended as a means of channeling discretion. *See* Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377, 404 (2006) (quoting Jody Freeman, *The Private Role in the Public Governance*, 75 N.Y.U. L. REV. 543, 664 (2000)); *see also* Jerry L. Mashaw, *Accountability and Institutional Design: Some Thoughts on the Grammar of Governance*, *in* PUBLIC ACCOUNTABILITY: DESIGNS, DILEMMAS, AND EXPERIENCES 115, 117 (Michael W. Dowdle ed., 2006). This is an admittedly nondemanding standard, but it is sufficient for present purposes because even on this understanding, content moderation systems are completely unaccountable.

[11] *See infra* section II.B, pp. 548–56.

leave up individual pieces of content (what this Article will call "paradigm cases"). It is a vast system of administration that includes a far broader range of decisions and decisionmakers than the standard picture admits.

This has become all the more true in recent years as platforms have started taking a more hands-on approach to content moderation in response to pressure from lawmakers and the public. "Content moderation," especially but not exclusively at the largest platforms, now includes many more things than it did even a few years ago: increased reliance on automated moderation; sticking labels on posts; partnerships with fact-checkers; greater platform and government collaboration; adding friction to how users share content; giving users affordances to control their own online experience; looking beyond the content of posts to how users *behave* online to determine what should be removed; and tinkering with the underlying dynamics of the very platforms themselves.[12] The people and processes that determine how user-generated content is treated on online platforms are therefore far more heterogeneous than depicted in the standard account. Content moderators include engineers, product managers, authorities outside platforms, teams monitoring behavioral signals, industry peers, and government partners. In short, content moderation is a complex and dynamic system, much of which will not be examined if the focus is on reviewing individual cases rather than their institutional context and interrelationships.

Second, the standard picture of content moderation leads regulators to assume that the primary way they can make social media platforms more publicly accountable is by requiring them to grant users ever more individual procedural rights. The allure of this assumption is understandable. Regulators turn to platform procedure because constitutional and practical limitations on governmental power mean that the job of setting most substantive content moderation rules cannot be taken away from private companies. Platforms can and will engage in content moderation beyond what the law could proscribe.[13] Platforms that removed only content that could be made illegal would rapidly become unusable, mired in spam, porn, harassment, and other graphic but not unlawful speech.[14] Many of the biggest content moderation controversies involve protected speech. In the United States, content like the Christchurch

[12] *See infra* section II.A, pp. 539–48.

[13] *See* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech* 13 (Hoover Working Grp. on Nat'l Sec., Tech. & L., Aegis Series Paper No. 1902, 2019).

[14] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 5 (2018).

Massacre livestream,[15] hate speech,[16] or coronavirus misinformation,[17] for example, cannot be legally proscribed. But it is a First Amendment right and a business interest for platforms to moderate.[18] Even if there were not constitutional obstacles to substantive governmental regulation of content moderation, the sheer scale, speed, and technological complexity of the task mean state actors could not directly commandeer the operations of content moderation. This is a descriptive, not normative, observation: the state simply does not have the capacity to usurp platforms as the frontline of content moderation.

For this reason, regulators are right to focus on procedure not substance. But the current regulatory focus on procedure in *individual cases* is a mistake. Precisely because so much of the work of content moderation occurs beyond the four corners of, and upstream of, individual cases, legislative procedural due process mandates for individual users can only do so much to improve the system — and, in some contexts, might harm a system's ability to achieve broader aims. More process is not always better process and maximizing the extent to which individuals feel they have been treated fairly is but one governance goal that content moderation should pursue. The scale and pace at which content moderation must operate make the tradeoffs between these individual interests and other goals such as overall speed, accuracy, and consistency especially acute.

This Article thus makes the case for a second-wave content moderation regulatory model: one based on a more comprehensive and complex view of content moderation systems than that presented by the standard picture. More specifically, it calls for an approach to content moderation regulation based on systems thinking, which focuses on the ex ante institutional design choices involved in creating a system of mass administration, rather than ex post individual error correction. Doing so makes it possible to imagine many more mechanisms of regulatory reform than the individual rights–focused model of content moderation suggested by a First Amendment analogy. The systems thinking approach draws instead on principles and practices of administrative law, which has long grappled with how to bring oversight and accountability to massive unelected bureaucracies at scale and in complex systems.[19]

[15] *See generally* Brandenburg v. Ohio, 395 U.S. 444 (1969) (per curiam) (establishing a very narrow test for when speech inciting unlawful action is unprotected by the First Amendment).

[16] *See generally* R.A.V. v. City of St. Paul, 505 U.S. 377 (1992) (holding that hate speech is protected by the First Amendment).

[17] *See generally* United States v. Alvarez, 567 U.S. 709 (2012) (holding that the government cannot proscribe speech purely because it is false).

[18] *See* Klonick, *supra* note 4, at 1626–30.

[19] *See* Jon D. Michaels, *An Enduring, Evolving Separation of Powers*, 115 COLUM. L. REV. 515, 532 (2015).

The project outlined in this Article is both ambitious and modest. The Article's reframing of content moderation governance is ambitious: it argues that most proposals for content moderation reform rely on an inaccurate understanding of the systems they seek to hold to account. The result is that what they achieve is accountability theater rather than accountability itself. This Article seeks to change that, by showing how lawmakers — once armed with a more accurate picture of how content moderation systems actually work — can regulate in ways that force platforms to be more accountable for the most consequential decisions they make.

This Article is modest insofar as it does not purport, or even attempt, to end all disputes about how to regulate online speech. It presumes that there is not — and never will be — agreement about what the problems with our current speech environment are or what the substantive rules for online speech should be. As such, the primary goal of content moderation governance should not be to resolve all substantive norms that govern online speech — an impossible and likely constitutionally prohibited ambition — but instead to design institutions that can provide the empirical foundations and channels for more productive disagreement.

The argument, therefore, is not that adopting a systems thinking approach would solve all content moderation problems for all time. Crucially, though, it would require an approach to regulation that is markedly different to many commonly proposed reforms, both in the United States and around the world.[20]

Taking this different approach would have many benefits that regulatory approaches based on the standard model would not have and avoid many counterproductive effects they would. Instead of fixing in place a uniform view of content moderation that is static and outdated, it would enable an iterative and dynamic approach that embraces the diversity and fluidity of content moderation. It would not be blind to, and therefore would not fail to make platforms accountable for, the important content moderation decisionmaking processes and institutions that fall out of the frame of the standard picture. It would force platforms to engage in dialogue about the value judgments that underpin the design and enforcement of their content moderation systems and for this reason make them accountable for (in the sense that they must reveal and explain) those judgments. And it would do so without requiring them to comply with government-dictated speech norms and so may

[20] This Article focuses on U.S. regulatory capacity but draws on examples from the global legal landscape to show the dominance of the standard picture and some attempts to move beyond it. The broader description of how content moderation works in practice, the pitfalls of regulating based on the standard picture, and the principles for a more productive approach this Article offers are, however, relevant to regulators everywhere.

be more politically feasible, and definitely more reconcilable with the First Amendment, than less systemic, more direct forms of online speech regulation. It would avoid creating an elaborate regulatory regime that creates barriers to entry and perverse incentives, and offers little in the way of effective remedy. And so it would avoid wasting monetary and political capital on ineffective reforms.

It would be comforting to think that content moderation could be made more accountable and effective if only platforms truly committed (or were forced to commit) to transposing the offline ex post review model of speech governance to the online world. Alas, such a simple fix is illusory. The task for reformers is much harder and requires embracing uncertainty and complexity in pursuit of a regulatory system that acknowledges the daunting legal, technical, and normative challenges that content moderation creates.

This Article takes up that task as follows. Part I describes the standard picture of content moderation arising from what this Article calls "the first wave" of content moderation literature that still dominates most discussion of online speech governance today. Part II then highlights that picture's blind spots: the systemic design choices that content moderation institutional designers confront in dealing with the speed and scale of online speech, and the heterogeneous institutions through which these choices are given effect.

Part III turns to the way the standard picture's stickiness in content moderation debates has led to misguided reform efforts. It shows that the individualistic and ex post mechanisms of accountability that the standard picture suggests, drawing from First Amendment analogies, will not surface or remedy systemic failures in content moderation systems. Many reform proposals rely on understandings of due process and transparency that are poorly suited to the goals that online speech governance should pursue.

Part IV outlines the project for a second wave of content moderation institutional design. It describes structural and procedural reforms that focus on ex ante and systemic accountability for the content moderation systems that determine the shape of the online public sphere. No regulatory approach can solve all of content moderation's current accountability deficits, but that is an illusory goal. What is needed is an experimental and incremental approach to content moderation governance that facilitates learning and iteration. In order to do this, regulators need an accurate understanding of the systems they seek to regulate. Regulation "cannot hope to promote systemic justice within a system that it fundamentally fails to comprehend."[21]

[21] Andrew Manuel Crespo, *Systemic Facts: Toward Institutional Awareness in Criminal Courts*, 129 HARV. L. REV. 2049, 2053 (2016).

I. THE STANDARD PICTURE OF CONTENT MODERATION

This Part sets out the standard picture of content moderation depicted in the first wave of content moderation scholarship, which dominates regulatory debates, especially in the United States.

This conception of "content moderation" is of a privatized hierarchical bureaucracy that applies legislative-style rules drafted by platform policymakers to individual cases and hears appeals from those decisions. A wealth of early and current academic, civil society, and public discourse about content moderation invokes this picture of content moderation.[22] This standard picture of content moderation revolves around paradigm cases involving "a platform's review of user-

[22] For an influential account of the standard picture, see Klonick, *supra* note 4, at 1639–41 (describing the three-tier structure of content moderation at Facebook). For further, by no means comprehensive, examples, see also GILLESPIE, *supra* note 14, at 116 ("[P]latforms currently impose moderation at scale by turning some or all users into an identification force, employing a small group of outsourced workers to do the bulk of the review, and retaining for platform management the power to set the terms."); Kyle Langvardt, *Can the First Amendment Scale?*, 1 J. FREE SPEECH L. 273, 298 (2021) ("Legal culture's reflexive answer to these kinds of problems . . . is to require 'some kind of a hearing.' The 'hearing' may include confrontation rights, protective burdens of proof and production, opportunities for appeal, and so on . . . . Many proposals to regulate or reform platform content moderation endorse this basic strategy, usually in combination with new transparency requirements." (footnotes omitted) (quoting Bd. of Regents v. Roth, 408 U.S. 564, 590 n.7 (1972))); Farzaneh Badiei et al., *Community Vitality as a Theory of Governance for Online Interaction*, 23 YALE J.L. & TECH. (SPECIAL ISSUE) 15, 33 (2021) ("[T]he major focus of many platform moderation efforts is simply to count and reduce individual violations."); 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* 21–22 (Hans-Bredow-Institut, Paper No. 1, 2020), https://leibniz-hbi.de/uploads/media/default/cms/media/oww9814_AP_WiP001InsideFacebook.pdf [https://perma.cc/W5B2-W5SD] (describing the "multi-step process" of rule development at Facebook, focusing on the Product Policy team which makes and changes the rules that are enforced by content moderators); ARTICLE 19, THE SOCIAL MEDIA COUNCILS: CONSULTATION PAPER 15 (2019), https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf [https://perma.cc/38QS-PLTA] (describing a proposal for a social media council that would sit above a platform's content moderation hierarchy and issue advisory opinions or an appeals mechanism in individual cases); DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 53–57 (2019) (describing a "mini-legislative session" the author attended, *id.* at 54, which had the "vibe of a law school seminar" and involved questions of notice, due process, and appeal that are "exactly the right questions you would hope Facebook would be asking itself," *id.* at 57); REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM 153–54 (2012) (describing the platform staff that develop policy and review procedures and "play the roles of lawmakers, judge, jury, and police all at the same time," *id.* at 154); Max Hoppenstedt, *A Visit to Facebook's Recently Opened Center for Deleting Content*, VICE (Jan. 2, 2018, 1:29 PM), https://www.vice.com/en/article/qv37dv/facebook-content-moderation-center [https://perma.cc/NG39-95Q7] (describing Facebook's processes for ensuring that all employees "interpret and apply the deletion rules in the same way" by applying the rules set by the Policy Team); Katrin Bennhold, *Germany Acts to Tame Facebook, Learning From Its Own History of Hate*, N.Y. TIMES (May 19, 2018), https://www.nytimes.com/2018/05/19/technology/facebook-deletion-center-germany.html [https://perma.cc/M3DQ-L3X4] ("Every day content moderators . . . pore over thousands of posts flagged by users

generated content posted on its site and the corresponding decision to keep it up or take it down."[23] Most high-profile content moderation controversies fall into this category. Think of Facebook's decision not to remove a doctored video of Speaker Nancy Pelosi that made her appear drunk,[24] platforms' treatment of President Trump's post stating "when the looting starts, the shooting starts" during 2020's Black Lives Matter protests,[25] platforms' varying responses to a *New York Post* article about Hunter Biden in the lead up to the 2020 U.S. election,[26] rules about Holocaust denial,[27] or how platforms will deal with twelve specific individuals allegedly responsible for the majority of online vaccine misinformation.[28] I could go on — the list of stories in this genre is essentially endless.

To tame this mind-boggling ocean of content posted on social media, large platforms adopt an "industrial" approach to content moderation,[29]

as upsetting or potentially illegal and make a judgment: Ignore, delete or, in particularly tricky cases, 'escalate' . . . ."); Marvin Ammori, *The "New"* New York Times*: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2276 (2014) ("[T]he terms of service function much as traditional laws do": as rules "to be operationalized by hundreds of employees and contractors around the world . . . ."); Jeffrey Rosen, *Google's Gatekeepers*, N.Y. TIMES (Nov. 28, 2008), https://www.nytimes.com/2008/11/30/magazine/30google-t.html [https://perma.cc/YT62-CZPD] ("Once flagged, a video is vetted by YouTube's internal reviewers at facilities around the world who decide whether to take it down, leave it up or send it up the YouTube hierarchy for more specialized review."); TIMOTHY GARTON ASH ET AL., GLASNOST! NINE WAYS FACEBOOK CAN MAKE ITSELF A BETTER FORUM FOR FREE SPEECH AND DEMOCRACY 9 (2019) ("Facebook has published some internal guidelines for the enforcement of Community Standards, and data about the enforcement of these standards, established an appeals process, and more than doubled the number of its content reviewers."); BEN BRADFORD ET AL., JUST. COLLABORATORY, YALE L. SCH., REPORT OF THE FACEBOOK DATA TRANSPARENCY ADVISORY GROUP 11–15 (2019), https://law.yale.edu/sites/default/files/area/center/justice/document/dtag_report_5.22.2019.pdf [https://perma.cc/48U4-SRZK] (outlining the Community Standards enforcement process, confined to takedowns and leave-ups and describing the appeals and review system).

[23] Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2427 (2020).

[24] Emily Stewart, *A Fake Viral Video Makes Nancy Pelosi Look Drunk. Facebook Won't Take It Down.*, VOX (May 24, 2019, 3:50 PM), https://www.vox.com/recode/2019/5/24/18638822/nancy-pelosi-doctored-video-drunk-facebook-trump [https://perma.cc/WHW3-JKT5].

[25] Cristiano Lima, *Zuckerberg: Facebook Leaving up Trump's "Shooting" Post*, POLITICO (May 29, 2020, 7:59 PM), https://www.politico.com/news/2020/05/29/zuckerberg-facebook-leaving-up-trumps-shooting-post-290292 [https://perma.cc/CK4R-Q77W].

[26] Adi Robertson, *Facebook and Twitter Are Restricting a Disputed* New York Post *Story About Joe Biden's Son*, THE VERGE (Oct. 14, 2020, 12:19 PM), https://www.theverge.com/2020/10/14/21515972/facebook-new-york-post-hunter-biden-story-fact-checking-reduced-distribution-election-misinformation [https://perma.cc/JPQ8-Z565].

[27] Jacob Kastrenakes, *Twitter Will Ban Holocaust Denial Posts, Following Facebook*, THE VERGE (Oct. 14, 2020, 4:21 PM), https://www.theverge.com/2020/10/14/21516468/twitter-holocaust-denial-banned-facebook-policy [https://perma.cc/G3X3-4Y7Y].

[28] Shannon Bond, *Just 12 People Are Behind Most Vaccine Hoaxes on Social Media, Research Shows*, NPR (May 14, 2021, 11:48 AM), https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes [https://perma.cc/4X4U-2JZA].

[29] GILLESPIE, *supra* note 14, at 77.

the goal of which is to create a "decision factory."[30] The frontline decisionmakers are the thousands of content moderators that review individual pieces of content and, increasingly, the automated tools that either flag pieces of content for human moderators to review or make moderation decisions without a human ever being involved.[31] The standard picture assumes that these frontline content moderator bureaucrats mindlessly (literally, in the case of automated tools)[32] implement the policy dictates they are given. Because mistakes (both human and machine) are inevitable, users can often (but not always)[33] appeal decisions made by this frontline of humans and artificial intelligence (AI), at which point the case will be subject to re-review.[34] The range of remedies is limited: the original decision is affirmed or reversed.

This picture was born of a first wave of content moderation scholarship that provided early insights into traditionally opaque systems, describing platforms as "New Governors" who performed content moderation much as judges do, applying their rules to pieces of content using something akin to legal reasoning.[35] The difficult and important decisions for platform policy teams in this account are in writing the substantive rules that determine the boundaries of free speech on social media.[36] Once those rules are written, it's simply a matter of applying them over and over . . . and over again — the standard picture conceives of content moderation as simply the aggregation of millions of daily paradigm cases.[37] The scale is hard to comprehend: in Q2 2022, Facebook took down 914,500,000 pieces of content,[38] YouTube took

[30] ROBYN CAPLAN, CONTENT OR CONTEXT MODERATION? ARTISANAL, COMMUNITY-RELIANT, AND INDUSTRIAL APPROACHES 23 (2018), https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf [https://perma.cc/K3CF-Q4NE].

[31] Elizabeth Dwoskin et al., *Content Moderators at YouTube, Facebook and Twitter See the Worst of the Web — and Suffer Silently*, WASH. POST (July 25, 2019, 1:00 AM), https://www.washingtonpost.com/technology/2019/07/25/social-media-companies-are-outsourcing-their-dirty-work-philippines-generation-workers-is-paying-price [https://perma.cc/KVQ4-2YFM].

[32] *See* Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41, 56 (2020); Robert Gorwa et al., *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, BIG DATA & SOC'Y, Jan.–June 2020, at 1, 4–5.

[33] *See* Klonick, *supra* note 4, at 1648.

[34] Tim Wu, *Will Artificial Intelligence Eat the Law? The Rise of Hybrid Social-Ordering Systems*, 119 COLUM. L. REV. 2001, 2016 (2019).

[35] Klonick, *supra* note 4, at 1642, 1663.

[36] *See* Kate Klonick, *Facebook v. Sullivan*, KNIGHT FIRST AMEND. INST. (Oct. 1, 2018), https://knightcolumbia.org/content/facebook-v-sullivan [https://perma.cc/JW49-44S5]; Monika Bickert, *Defining the Boundaries of Free Speech on Social Media*, *in* THE FREE SPEECH CENTURY 254 (Lee C. Bollinger & Geoffrey R. Stone eds., 2019); 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 [https://perma.cc/D6WR-64A7].

[37] *See* Klonick, *supra* note 23, at 2432–33.

[38] *See* FACEBOOK, COMMUNITY STANDARDS ENFORCEMENT REPORT (2022).

down 3,987,509 channels and 4,496,933 videos,[39] and in Q1 2022, TikTok removed 102,305,516 videos.[40] These figures do not include every time these platforms decided to *leave up* content flagged for review (which would greatly exceed decisions to remove content) or appeals. Smaller platforms deal with smaller numbers, but with fewer resources and less technical capacity.

The picture of content moderation the standard model suggests is, in other words, something like a caricature of the Weberian model of bureaucracy: a bureaucratic organization as a transmission belt implementing rules in an efficient and reliable way, organized around a limited set of hierarchically organized institutions and rights of appeal.[41]

Which is to say, contemporary discussion of platform regulation focuses on paradigm cases — those that evoke the "day-in-court ideal",[42] and most closely look like famous traditional free speech cases in which individual utterances get carefully measured against lofty speech rules and principles. This view of content moderation naturally invokes analogies to the practice of offline constitutional law. If the work of content moderation primarily involves the application of specific speech rules to particular cases within predetermined categories of content, then the questions it raises resemble those raised in First Amendment cases. Much of this literature is not normative: it does not argue that content moderation *should* resemble the standard picture. Instead, it is descriptive — assuming that this *is* how content moderation works, and then discussing reforms that might improve the functioning of that system, which it takes as a given.

The standard picture is not wrong per se: there are indeed policy teams within platforms making up rules and large bureaucracies of content moderators applying them. But the standard picture *is* outdated and incomplete, and decisions in paradigm cases are downstream of more consequential choices about institutional and platform design. Basing regulation around the standard picture of content moderation threatens to, at best, regulate the content moderation systems of 2016, not the ones that exist now. At worst, it entrenches dominant firms

[39] *See YouTube Community Guidelines Enforcement*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/youtube-policy/removals?hl=en_GB [https://perma.cc/9XPN-N952].

[40] TIKTOK, COMMUNITY GUIDELINES ENFORCEMENT REPORT (2022), https://www.tiktok.com/transparency/en-us/community-guidelines-enforcement-2022-1 [https://perma.cc/3S8C-N6GA].

[41] MAX WEBER, ECONOMY AND SOCIETY: AN OUTLINE OF INTERPRETIVE SOCIOLOGY 957 (Guenther Roth & Claus Wittich eds., 1978) (describing bureaucracy as involving "a clearly established system of super- and subordination in which there is a supervision of the lower offices by higher ones," as well as "the possibility of appealing, in a precisely regulated manner, the decision of a lower office to the corresponding superior authority").

[42] Martin H. Redish & Julie M. Karaba, *One Size Doesn't Fit All: Multidistrict Litigation, Due Process, and the Dangers of Procedural Collectivism*, 95 B.U. L. REV. 109, 133–35 (2015).

without meaningfully changing their operations by basing regulation around a limited slice of what they already do.

II. THE STANDARD PICTURE'S BLIND SPOTS

This Part analyzes the blind spots of the standard picture, showing how it rests on flawed descriptive and theoretical foundations. Descriptively, it is blind to two important characteristics: the wide diversity of institutions involved in content moderation outside the hierarchical bureaucracy that is the content moderation appeals system, and the wide variety of ex ante tradeoffs that content moderation institutional designers have to engage with. Theoretically, it assumes the necessity of a model of speech governance and the judicial role adapted from the First Amendment context that is intuitive but ill-suited to the scale and speed at which content moderation governance must operate.

A. Content Moderation Bureaucracies Are a "They" Not an "It"

Content moderation is far more than just the paradigm cases. Some of the most important decisions about how user-generated content is treated on platforms are made outside the borders of the standard picture. Content moderation bureaucracies are a "they" not an "it": they are made up of a sprawling array of actors and institutions, each of which has different functions and goals.[43] This section provides a map for the heterogeneous institutions involved in content moderation, which are normally studied in isolation (and often not described as content moderation at all, despite involving decisions about the content on platforms). This is not intended to be a comprehensive description of any single platform, let alone all platforms — indeed, the key takeaway should be that content moderation defies one-size-fits-all understanding. What these actors have in common is that they all play large roles in determining the shape of the online speech environment but disappear from view when one looks at content moderation purely through the frame of the standard picture.

1. Non-Content-Based Content Moderation. — The standard picture of content moderation does not account for the increasing number of interventions platforms make on the basis of considerations other than the *content* of posts. Platforms increasingly intervene based on the

[43] *Cf.* Kenneth A. Shepsle, *Congress Is a "They," Not an "It": Legislative Intent as Oxymoron*, 12 INT'L REV. L. & ECON. 239 (1992); Elizabeth Magill & Adrian Vermeule, *Allocating Power Within Agencies*, 120 YALE L.J. 1032, 1036 (2011) (extending the congressional analogy to administrative agencies).

behavior of groups of accounts and the *actors* or *associations* behind them.[44]

Platforms' moderation of influence or information operations is an example of this kind of behavioral content moderation (that is, content moderation based on how users *behave*, not what they post).[45] Indeed, it was the need to create a policy to deal with the ways in which Russian actors exploited social media during the 2016 U.S. presidential election that led to the first behavioral content moderation policies.[46] How exactly platforms determine when users cross the line into illegitimate platform manipulation is unclear: behavioral content moderation is extremely opaque,[47] and this is in part by design — platforms insist that transparency about these policies would only enable bad actors to evade detection.[48] This is the first difference between the paradigm case and behavioral content moderation: giving notice and reasons to users is seen as undermining the effectiveness of rules rather than promoting compliance.[49]

A second important difference is that when it comes to behavioral content moderation, any individual post, standing alone, may not appear (or, indeed, be) problematic, but a group of accounts as a whole

[44] Camille François, *Actors, Behaviors, Content: A Disinformation ABC* 2 (Transatlantic High Level Working Grp. on Content Moderation Online and Freedom of Expression, 2019), https://www.ivir.nl/publicaties/download/ABC_Framework_2019_Sept_2019.pdf [https://perma.cc/4EZP-W3TK]; Camille François & Evelyn Douek, *The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting About Information Operations*, 1 J. ONLINE TRUST & SAFETY 1, 1–2 (2021).

[45] The definitions of "information operations" and "influence operations" are notoriously slippery, often opaque, and differ from platform to platform; for a related discussion, see François & Douek, *supra* note 44, at 11. Here I am not using the term technically, and just mean to refer generally to concerted and coordinated efforts done with the intention of influencing others' beliefs or opinions.

[46] *Id.* at 7.

[47] Evelyn Douek, *The Free Speech Blind Spot: Foreign Election Interference on Social Media*, *in* DEFENDING DEMOCRACIES: COMBATING FOREIGN ELECTION INTERFERENCE IN A DIGITAL AGE 265, 270–72 (Duncan B. Hollis & Jens David Ohlin eds., 2021); Evelyn Douek, *What Does "Coordinated Inauthentic Behavior" Actually Mean?*, SLATE (July 2, 2020, 5:26 PM), https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html [https://perma.cc/Q3FS-JF7Y].

[48] *See, e.g.*, GOOGLE, HOW GOOGLE FIGHTS DISINFORMATION 3 (2019), https://www.blog.google/documents/37/How_Google_Fights_Disinformation.pdf [https://perma.cc/3MEB-VRTR] ("[W]e try to be clear and predictable in our efforts, letting users and content creators decide for themselves whether we are operating fairly. Of course, this is a delicate balance, as sharing too much of the granular details of how our algorithms and processes work would make it easier for bad actors to exploit them."); Sara Harrison, *Twitter's Disinformation Data Dumps Are Helpful — To a Point*, WIRED (July 7, 2019, 7:00 AM), https://www.wired.com/story/twitters-disinformation-data-dumps-helpful [https://perma.cc/UJT2-QKYB] ("Twitter would not reveal any specifics about its process for this article. 'We seek to protect the integrity of our efforts and avoid giving bad actors too much information, but in general, we focus on conduct, rather than content.'").

[49] *Cf.* Badiei et al., *supra* note 22, at 42.

may violate the rules.[50] Many of the posts found to be connected to the Russian campaigns targeting the 2016 U.S. election fell into this category. As such, looking at an individual case cannot show if a behavioral content moderation decision was correct.

Third, behavioral content moderation is often done by teams separate from those in the standard picture. These takedowns are typically handled by cybersecurity-oriented teams, like Google's "Threat Analysis Group"[51] or Twitter's "Site Integrity" team,[52] which are distinct from the policy teams that handle platforms' rules more generally. Reflecting this distinction, behavioral content moderation is often reported in separate documents when platforms release voluntary transparency reports about their enforcement actions.[53]

Information operations are the highest profile example of behavioral content moderation, but by far the largest categories of content takedowns — spam and fake accounts — are also done on the basis of behavioral signals. Spam, which means different things to different platforms,[54] gets almost no attention in content moderation debates and it's unclear why not. "Spam" is not an objective category, but a construct implicating value judgments about acceptable and unacceptable online conduct.[55] Early on, there were concerns about the free speech implications of the inevitable false positives caught in spam filters,[56] but the practical necessity of spam filtering means it is now an uncontroversial and "essential part of the internet."[57] Indeed, "[r]emoving spam is censoring content; it just happens to be content that nearly all users agree should go."[58]

Behavioral content moderation is becoming increasingly important as platforms expand their policies to include conspiracy groups and behaviors such as coordinated harassment, mass-coordinated reporting,

[50] *See* RENEE DIRESTA ET AL., THE TACTICS & TROPES OF THE INTERNET RESEARCH AGENCY 99 (2019), https://digitalcommons.unl.edu/senatedocs/2 [https://perma.cc/K5YF-K9TR].

[51] *Threat Analysis Group (TAG)*, BLOG.GOOGLE, https://blog.google/threat-analysis-group [https://perma.cc/42GJ-7V56].

[52] *Twitter's Head of Site Integrity, On Fighting Election Disinformation*, NPR (Mar. 2, 2020, 4:18 PM), https://www.npr.org/2020/03/02/811338220/twitter-head-of-site-integrity-on-fighting-election-disinformation [https://perma.cc/4WZD-2XLY].

[53] François & Douek, *supra* note 44, at 9.

[54] Graggle, *How We're Fighting Spammers on Discord*, DISCORD BLOG (Nov. 12, 2021), https://discord.com/blog/how-discord-is-fighting-spam [https://perma.cc/N3FN-7DQZ] ("Definitions on what 'spam' is can vary widely across companies . . . .").

[55] FINN BRUNTON, SPAM: A SHADOW HISTORY OF THE INTERNET xiv–xv (2013).

[56] *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/2R6H-X9L7] ("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.").

[57] SARAH JEONG, THE INTERNET OF GARBAGE 67 (2018).

[58] GILLESPIE, *supra* note 14, at 217 n.22.

and brigading, taking down tens of thousands of accounts under these umbrellas with little explanation.[59] The borders of these categories are ambiguous. The ambit of these new policies, like Twitter's "coordinated harmful activity" policy[60] and Facebook's "coordinated social harm" policy,[61] remains completely unclear. When does a deluge of tweets become coordinated harassment? How does a platform determine if an individual user is coordinating with others? When does the way a group acts online or offline create "social harm"? These questions won't be answered by looking at a single piece of content as in a paradigm case. After a coauthor and I argued that the opacity of these distinctions was problematic, Twitter expanded its previously sui generis data transparency for information operations to these categories as well, acknowledging that there is no good reason for them to be treated differently.[62] Other platforms have yet to follow.

In short, these categories of content moderation are frequently managed by a completely different part of a platform's bureaucracy and according to different procedures and rules to those depicted in the standard picture. Often, much of this activity does not even get called "content moderation" even though it involves the removal of content and accounts based on a platform's terms of service. For this reason, perhaps, most regulatory proposals have little to say about this kind of platform action. This is a mistake: all the same accountability deficits apply, if not more so.

2. *Cross-Platform and Government Cooperation.* — The standard picture depicts platforms as standalone institutions and does not account for the extent to which platforms increasingly cooperate with each other and with governments.[63] This kind of collaboration has dramatically increased in the last half decade. To moderate child sexual abuse material (CSAM) and terrorist and extremist content, many platforms share

[59] *An Update to How We Address Movements and Organizations Tied to Violence*, FACEBOOK NEWSROOM (Nov. 9, 2021, 10:00 AM), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence [https://perma.cc/9WSH-WFGZ] (originally published Aug. 19, 2020); Evelyn Douek, *Twitter Brings Down the Banhammer on QAnon*, LAWFARE (July 24, 2020, 2:56 PM), https://www.lawfareblog.com/twitter-brings-down-banhammer-qanon [https://perma.cc/7Q8B-AZUE].

[60] *Coordinated Harmful Activity*, TWITTER HELP CTR. (Jan. 2021), https://help.twitter.com/en/rules-and-policies/coordinated-harmful-activity [https://perma.cc/2X8M-UDF9].

[61] Nathaniel Gleicher, *Removing New Types of Harmful Networks*, FACEBOOK NEWSROOM (Sept. 16, 2021), https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks [https://perma.cc/QN2Z-ELKX].

[62] François & Douek, *supra* note 44; Yoel Roth & Vijaya Gadde, *Expanding Access Beyond Information Operations*, TWITTER BLOG (Aug. 24, 2022), https://blog.twitter.com/en_us/topics/company/2021/-expanding-access-beyond-information-operations- [https://perma.cc/X4YF-3DL7] (originally published Dec. 2, 2021).

[63] Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INST. (Feb. 11, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels [https://perma.cc/W333-KC8F].

common databases of violating material.[64] To detect influence operations, platforms rely on tips from one another, governments, and third-party analysts.[65] Platforms insist that their decisionmaking remains independent,[66] perhaps to avoid political or antitrust scrutiny, but the *whole point* of such collaborations is that each of the parties informs the decisionmaking of the others; to what extent is unknown.

Even though this kind of content moderation occurs on the basis of collaborative arrangements, mechanisms to create accountability for such arrangements and remedies for mistakes remain individualized and reactive. If content is mistakenly taken down across multiple platforms as the result of collective action, the user may not even be aware of that fact and will only be able to challenge the decision via each platform individually. From the outside, each decision looks like any other individual takedown. But in reality, the institutional context, dynamics, and processes are very different and the biases of one party may influence the others. Content moderation literature has rightly raised concerns about the ways in which governments can influence paradigm cases to perform censorship beyond what the state could compel by law (known as "jawboning").[67] What the standard picture misses are instances when that influence is not viewed as illegitimate pressure but an intentional part of institutional design.[68]

3. Delegated Decisionmaking. — Increasingly, platforms outsource judgments to third parties in order to disavow responsibility for outcomes in the paradigm cases contemplated by the standard picture.[69]

[64] *Id.*

[65] *See, e.g.*, THE ELECTION INTEGRITY P'SHIP, THE LONG FUSE: MISINFORMATION AND THE 2020 ELECTION 9–10 (2021).

[66] Issie Lapowsky, *This Big Tech Group Tried to Redefine Extremism. It Got Messy.*, PROTOCOL (July 26, 2021), https://www.protocol.com/policy/gifct-erin-saltman [https://perma.cc/9USF-FWNG].

[67] *See, e.g.*, Derek E. Bambauer, *Against Jawboning*, 100 MINN. L. REV. 51, 60 (2015); Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114 (2018); Genevieve Lakier, *Informal Government Coercion and the Problem of "Jawboning*,*"* LAWFARE (July 26, 2021, 3:52 PM), https://www.lawfareblog.com/informal-government-coercion-and-problem-jawboning [https://perma.cc/C959-9ZJB].

[68] *See, e.g.*, Elena Chachko, *National Security by Platform*, 25 STAN. TECH. L. REV. 55, 86–90 (2021). For a high-level exploration of some of the risks involved in such platform-government collaborations, see BSR, HUMAN RIGHTS ASSESSMENT: GLOBAL INTERNET FORUM TO COUNTER TERRORISM (2021), https://gifct.org/wp-content/uploads/2021/07/BSR_GIFCT_HRIA.pdf [https://perma.cc/P3FH-7C49].

[69] The Oversight Board is distinct from the institutions described in this section, which are involved in the *initial* content moderation decision, whereas the Board is intended to be an independent appeals mechanism.

Fact-checking partnerships are the most significant such relationships. Platforms long insisted they should not be "arbiters of truth,"[70] and refused to take content down based on falsity alone. But pressure from lawmakers and the public to *do something* about the "age of disinformation"[71] made a completely hands-off approach politically and commercially untenable. Relying on third-party fact-checkers' judgments to guide content moderation has been a favored solution because it appears to preserve platforms' neutrality while also responding to concerns about the spread of false information.[72]

But the devil is in the details. Platforms decide which fact-checkers to trust, what content fact-checkers will have access to, and how fact-checking feeds into platform decisionmaking. There are few details about how platforms' fact-checking programs work and the way these programs are structured can be used to mask substantive choices. For example, reporting indicates that Facebook can remove stories from the fact-checking pool by labelling them "opinion,"[73] or can "remove strikes" from repeat-offenders' accounts so they avoid the typical penalties such as demotion in search results.[74] TikTok touts its fact-checking program but doesn't provide any details. There is no public information about how often or why the platform refers content to fact-checkers or how

[70] *See, e.g.*, Mark Zuckerberg, FACEBOOK (Nov. 19, 2016), https://www.facebook.com/zuck/posts/10103269806149061 [https://perma.cc/2TVL-4Z4B]; Callum Borchers, *Twitter Executive on Fake News: "We Are Not the Arbiters of Truth*,*"* WASH. POST (Feb. 8, 2018, 3:20 PM), https://www.washingtonpost.com/news/the-fix/wp/2018/02/08/twitter-executive-on-fake-news-we-are-not-the-arbiters-of-truth [https://perma.cc/9UMU-DDXZ]; Supraja Srinivasan, *We Don't Want to Be Arbiters of Truth: YouTube CBO Robert Kyncl*, ECON. TIMES (Mar. 24, 2018, 8:52 AM), https://tech.economictimes.indiatimes.com/news/internet/we-dont-want-to-be-arbiters-of-truth-youtube-cbo-robert-kyncl/63438805 [https://perma.cc/L7JK-5EEC].

[71] *See, e.g.*, Emily Bazelon, *Free Speech Will Save Our Democracy*, N.Y. TIMES MAG. (Oct. 13, 2020), https://www.nytimes.com/2020/10/13/magazine/free-speech.html [https://perma.cc/2Z4N-HKRZ]; THE DISINFORMATION AGE: POLITICS, TECHNOLOGY, AND DISRUPTIVE COMMUNICATION IN THE UNITED STATES (W. Lance Bennett & Steven Livingston eds., 2020).

[72] *See Fact Checks in YouTube Search Results*, YOUTUBE HELP, https://support.google.com/youtube/answer/9229632?hl=en [https://perma.cc/P2T7-3LHA]; Arjun Narayan Bettadapur, *TikTok Partners with Fact-Checking Experts to Combat Misinformation*, TIKTOK NEWSROOM (Oct. 1, 2020), https://newsroom.tiktok.com/en-au/tiktok-partners-with-fact-checking-experts-to-combat-misinformation [https://perma.cc/5ZHK-VBA3]; *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/R8TK-TQCQ].

[73] Veronica Penney, *How Facebook Handles Climate Disinformation*, N.Y. TIMES (Sept. 14, 2020, 6:00 AM), https://www.nytimes.com/2020/07/14/climate/climate-facebook-fact-checking.html [https://perma.cc/XWP3-WNAQ] (originally published July 14, 2020).

[74] Isaac Stanley-Becker & Elizabeth Dwoskin, *Trump Allies, Largely Unconstrained by Facebook's Rules Against Repeated Falsehoods, Cement Pre-election Dominance*, WASH. POST (Nov. 1, 2020), https://www.washingtonpost.com/technology/2020/11/01/facebook-election-misinformation [https://perma.cc/P76F-Z8Y6].

fact-checks are used.[75] Like Facebook, TikTok attempts to harness the legitimacy dividends of working with outside experts while structuring the relationship in a way that allows it to retain discretion. In these ways, arrangements designed to ensure the appearance of platform neutrality are plagued with procedural loopholes that subvert that very purpose.[76]

Content moderation during the pandemic and platforms' reliance on "authoritative sources" like the World Health Organization to determine what constituted coronavirus misinformation are other examples of delegated decisionmaking.[77] In hard cases, such as where designated authorities became increasingly out of step with scientific consensus (for example, guidance on mask wearing early in the pandemic), lack of clarity around the limits of those authorities' influence on content moderation can become controversial.[78]

4. Design and Affordances. — Focusing on individual cases of content moderation failure also ignores the ex ante design choices platforms make from which all paradigm cases are downstream.[79]

These design choices are some of the most important in content moderation. Don't take my word for it — platforms agree. Facebook CEO Mark Zuckerberg, in a blog post titled *A Blueprint for Content Governance*, described how, no matter where you draw a policy line, content that approaches that line will get more engagement, so rather than simply removing content, the more effective way of dealing with misinformation is to "reduc[e] its distribution and virality."[80] Facebook's

[75] Bettadapur, *supra* note 72; Sarah Perez, *TikTok to Flag and Downrank "Unsubstantiated" Claims Fact Checkers Can't Verify*, TECHCRUNCH (Feb. 3, 2021, 8:00 AM), https://techcrunch.com/2021/02/03/tiktok-to-flag-and-downrank-unsubstantiated-claims-fact-checkers-cant-verify [https://perma.cc/UE8R-CMXW].

[76] Mike Ananny, *The Partnership Press: Lessons for Platform-Publisher Collaborations as Facebook and News Outlets Team to Fight Misinformation*, COLUM. JOURNALISM REV. (Apr. 4, 2018), https://www.cjr.org/tow_center_reports/partnership-press-facebook-news-outlets-team-fight-misinformation.php [https://perma.cc/6U43-442R] ("Facebook can thus claim that it does not make final judgments about which content is fake or real — fact-checkers ostensibly are the experts in judging the character of dashboard stories — but through a complex and proprietary set of classifications, it powerfully sets the conditions that surface and characterize misinformation.").

[77] Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV. 759, 832 (2021).

[78] Evelyn Douek, *More Content Moderation Is Not Always Better*, WIRED (June 2, 2021, 8:00 AM), https://www.wired.com/story/more-content-moderation-not-always-better [https://perma.cc/NYN7-UGGL].

[79] *See* Ari Ezra Waldman, *Disorderly Content*, 97 WASH. U. L. REV. (forthcoming 2022) (manuscript at 16–17) (on file with the Harvard Law School Library), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3906001 [https://perma.cc/U4S6-XYZV] (noting the current literature's failure to examine this aspect of content moderation).

[80] Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (May 5, 2021), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 [https://perma.cc/X3UJ-EB8K] (originally published Nov. 15, 2018).

transparency reports on its content moderation highlight these measures: in boasting about reductions in how often users saw hate speech, Facebook said the main reason was not better rules, or better moderation tools, or more human moderators, but changes to the News Feed to reduce "the number of times we display posts that later may be determined to violate our policies."[81] The same is true of YouTube: it touts progress in combatting violating content on its site by adjusting its recommendation algorithm to demote "borderline" content.[82] YouTube's chief product officer has said design choices play an "order or two orders of magnitude more impactful role" than simply removing content.[83]

These changes are not "content moderation" in the sense that the paradigm case contemplates; they are changes to what Professor Julie Cohen has called content *immoderation*.[84] Platforms can control the amount of violating content users see not by removing more or fewer posts but by changing those posts' distribution. Yet content immoderation surely involves the same false positives and negatives as content moderation,[85] and effects of demotions might be as significant as removals in terms of whether a piece of content reaches an audience. However, a user may not be aware that a platform made any intervention if the content is not removed.

The possibilities for content moderation through platform design interventions are limited more by engineers' imaginations than any fixed technological fact. Two trends are worth highlighting.

First, in the past year platforms have started introducing nudges designed to minimize the amount of violating content users post or share in the first place.[86] Examples include prompts to think twice before posting certain types of content ("[W]ould you like to reconsider posting

[81] Guy Rosen, *Community Standards Enforcement Report, Fourth Quarter 2020*, FACEBOOK NEWSROOM (Feb. 11, 2021), https://about.fb.com/news/2021/02/community-standards-enforcement-report-q4-2020 [https://perma.cc/BAW7-2GNR].

[82] Julia Alexander, *YouTube Claims Its Crackdown on Borderline Content Is Actually Working*, THE VERGE (Dec. 3, 2019, 12:00 PM), https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources [https://perma.cc/D2BF-HDJL].

[83] Nilay Patel, *YouTube Chief Product Officer Neal Mohan on the Algorithm, Monetization, and the Future for Creators*, THE VERGE (Aug. 3, 2021, 12:00 PM), https://www.theverge.com/22606296/youtube-shorts-fund-neal-mohan-decoder-interview [https://perma.cc/W39N-LZHH].

[84] JULIE E. COHEN, BETWEEN TRUTH AND POWER 100, 136 (2019).

[85] *See* Craig Timberg et al., *Inside Facebook, Jan. 6 Violence Fueled Anger, Regret over Missed Warning Signs*, WASH. POST (Oct. 22, 2021, 7:36 PM), https://www.washingtonpost.com/technology/2021/10/22/jan-6-capitol-riot-facebook [https://perma.cc/6P7K-UYHS] ("Facebook has declined to deploy some mitigation tactics when chief executive Mark Zuckerberg has objected on the basis that they will cause too many 'false positives' or might stop people from engaging with its platforms.").

[86] *See* Matthew Katsaros et al., *Reconsidering Tweets: Intervening During Tweet Creation Decreases Offensive Content*, INT'L CONF. ON WEBLOGS & SOC. MEDIA (May 26, 2020) ("More recently, platforms have begun employing moderation approaches which seek to intervene prior to offensive content being posted.").

this? This comment contains words that may violate our Community Guidelines")[87] and adding "friction," such as extra clicks or limits on how many times a message can be forwarded.[88] Minor interventions can have large effects: a simple pop-up from Twitter asking users if they wanted to read an article before retweeting it prompted 40% more users to open the article.[89] When WhatsApp reduced the number of times users could forward a message at once, forwards decreased by 25%.[90]

The second trend in platform design has been to give users more control. This includes measures like allowing users themselves to control who can engage with their content,[91] create preferences for what kind of content gets recommended to them,[92] mass-delete comments on their posts,[93] or designate their own forum administrators or "experts."[94]

In sum, content moderation interventions are heterogenous and will no doubt be even more so by the time this Article is published. Indeed, platforms are increasingly turning to the interventions described here precisely *because* they are not paradigm cases and *because* they are outside the standard picture. Such interventions are less controversial and attract less attention because they are seen to be content-neutral and do not require platforms to make decisions to remove individual posts. But

[87] Tara Wadhwa, *New Tools to Promote Kindness on TikTok*, TIKTOK NEWSROOM (Mar. 10, 2021), https://newsroom.tiktok.com/en-us/new-tools-to-promote-kindness [https://perma.cc/BDR7-RL9M]; *see also* Anita Patwardhan Butler & Alberto Parrella, *Tweeting with Consideration*, TWITTER BLOG (May 5, 2021), https://blog.twitter.com/en_us/topics/product/2021/tweeting-with-consideration [https://perma.cc/CX48-NS4J]; Sarah (TeamYouTube), *Help Us Keep Comments Respectful — New Community Guidelines Comment Reminders*, YOUTUBE HELP (Dec. 3, 2020), https://support.google.com/youtube/thread/86685658/help-us-keep-comments-respectful-%E2%80%93-new-community-guidelines-comment-reminders [https://perma.cc/9KQK-RWVA].

[88] *See, e.g.*, Alex Kantrowitz, *Facebook and Twitter Are Rethinking Their Share Buttons*, MEDIUM (Oct. 30, 2020), https://onezero.medium.com/facebook-and-twitter-are-rethinking-their-share-buttons-32ad01ed1bfc [https://perma.cc/3N7E-3XFT]; Jon Porter, *WhatsApp Says Its Forwarding Limits Have Cut the Spread of Viral Messages by 70 Percent*, THE VERGE (Apr. 27, 2020, 8:28 AM), https://www.theverge.com/2020/4/27/21238082/whatsapp-forward-message-limits-viral-misinformation-decline [https://perma.cc/4SGV-4QWS].

[89] @TwitterComms, TWITTER (Sept. 24, 2020, 1:11 PM), https://twitter.com/TwitterComms/status/1309178716988354561 [https://perma.cc/GQD8-C6F9].

[90] *Keeping WhatsApp Personal and Private*, WHATSAPP BLOG (Apr. 7, 2020), https://blog.whatsapp.com/Keeping-WhatsApp-Personal-and-Private/?lang=en [https://perma.cc/28KP-MDF5].

[91] *See, e.g.*, Alex Hern, *Social Network Giants Pledge to Tackle Abuse of Women Online*, THE GUARDIAN (July 1, 2021, 1:00 PM), http://www.theguardian.com/society/2021/jul/01/social-networks-facebook-google-twitter-tiktok-pledge-to-tackle-abuse-of-women-online [https://perma.cc/G22M-WVJE].

[92] *See, e.g.*, *Introducing Sensitive Content Control*, INSTAGRAM BLOG (July 20, 2021), https://about.instagram.com/blog/announcements/introducing-sensitive-content-control [https://perma.cc/3XTW-WCUJ].

[93] *See, e.g.*, Joshua Goodman, *New Tools to Combat Bullying on TikTok*, TIKTOK NEWSROOM (May 20, 2021), https://newsroom.tiktok.com/en-us/new-tools-to-combat-bullying [https://perma.cc/M3VP-4ADY].

[94] *See, e.g.*, Maria Angelidou-Smith, *New Ways to Elevate Knowledgeable Experts in Facebook Groups*, FACEBOOK NEWSROOM (July 13, 2021), https://about.fb.com/news/2021/07/new-ways-to-elevate-experts-in-facebook-groups [https://perma.cc/33FX-4HJY].

these different forms of content moderation all exhibit the same accountability deficits, the same information asymmetries, and the same capacity to influence the online information ecosystem. Regulation that fixates on the singular, narrow model of content moderation represented by paradigm cases in the standard picture risks overlooking many of the most important forms of platform decisionmaking, becoming outdated before it is enacted, and locking in a form of oversight that is limited in its ambition.

B. Content Moderation Is All About Tradeoffs

In focusing on how platforms apply their rules in individual cases, the standard picture ignores that content moderation systems must pursue multiple governance goals beyond single-mindedly arriving at "correct" decisions in accordance with their rules. As with all regulatory institutions, platforms face competing demands for efficiency, accuracy, responsiveness to stakeholders, and commitment to procedural rule-of-law values.[95] Often, these values will be in tension. The way tradeoffs are resolved reflects substantive value judgments that are embedded into content moderation systems at a point upstream from paradigm cases. Once a paradigm case presents itself, these value judgments have already determined the universe of possible outcomes.

1. Error Choices: False Positives Versus False Negatives. — Given the unfathomable scale of content moderation, errors are inevitable.[96] Accepting and choosing between errors is therefore a central consideration in content moderation institutional design.[97] Error minimization at all costs may not always be a desirable goal, much less a realistic one, and in many cases decisionmakers may choose to err on the side of more errors rather than less.

[95] *See* Jerry L. Mashaw, *Structuring a "Dense Complexity": Accountability and the Project of Administrative Law*, 5 ISSUES IN LEGAL SCHOLARSHIP 1, 13–14 (2005) (citing Gunther Teubner, *Juridification: Concepts, Aspects, Limits, Solutions*, *in* JURIDIFICATION OF SOCIAL SPHERES: A COMPARATIVE ANALYSIS IN THE AREAS OF LABOR, CORPORATE, ANTITRUST AND SOCIAL WELFARE LAW 3 (Gunther Teubner ed., 1987); MICHAEL ASIMOW, FEDERAL ADMINISTRATIVE ADJUDICATION OUTSIDE THE ADMINISTRATIVE PROCEDURE ACT 59–60 (2019), https://www.acus.gov/sites/default/files/documents/Federal%20Administrative%20Adj%20Outside%20the%20APA%20-%20Final.pdf [https://perma.cc/CZ6E-MBSH].

[96] *See* Mike Masnick, *Masnick's Impossibility Theorem: Content Moderation at Scale Is Impossible to Do Well*, TECHDIRT (Nov. 20, 2019, 9:31 AM), https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml [https://perma.cc/RD4K-WKGY]; James Grimmelman, *To Err Is Platform*, KNIGHT FIRST AMEND. INST. (Apr. 6, 2018), https://knightcolumbia.org/content/err-platform [https://perma.cc/4CPG-2JHY]; Douek, *supra* note 77, at 792.

[97] 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/N52Y-4VVM].

The most obvious countervailing consideration is timeliness. The standard picture ignores the fact that "wrong" decisions on the substance may be the product of a (reasonable) ex ante decision to prioritize speed. Offline speech disputes in courts can take months if not years to be resolved[98] and even then there are persistent disagreements about whether courts got it "right." Online timescales are entirely different. Even slightly delayed "right" decisions can be irrelevant decisions.[99] Posts can go viral and cause harm within minutes. To make decisions that are even remotely timely, in many cases the largest platforms rely on automated tools that are insensitive to context and provide little explanation for their decisions.[100] The livestream of the Christchurch massacre was online for less than seventy minutes but spread far and wide.[101] In attempting to scrub the video from their services quickly, platforms consciously made the "wrong" decision to take down news reports featuring portions of the video.[102] Lawmakers rarely engage with this tradeoff between careful consideration of each case and the need for content moderation to be quick in order to be effective: laws frequently impose ever-shorter deadlines for platforms to take content down while also expecting context-specific adjudications and procedural protections for users.[103]

Content moderation system designers might also reasonably choose to err on the side of more false positives in high-risk situations where the costs of false negatives increase. Major platforms' decisions to err on the side of over-removal of misinformation during the COVID-19 pandemic because the costs of under-removal during a public health

[98] Jacob Mchangama, *Rushing to Judgment: Examining Government Mandated Content Moderation*, LAWFARE (Jan. 26, 2021, 8:00 AM), https://www.lawfareblog.com/rushing-judgment-examining-government-mandated-content-moderation [https://perma.cc/4GJ8-PXTX].

[99] David Kaye (Special Rapporteur), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, at 13, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) ("Even with appeal, however, the remedies available to users appear limited or untimely to the point of non-existence . . . .").

[100] CAREY SHENKMAN ET AL., DO YOU SEE WHAT I SEE? CAPABILITIES AND LIMITS OF AUTOMATED MULTIMEDIA CONTENT ANALYSIS 8 (2021), https://cdt.org/wp-content/uploads/2021/05/2021-05-18-Do-You-See-What-I-See-Capabilities-Limits-of-Automated-Multimedia-Content-Analysis-Full-Report-2033-FINAL.pdf [https://perma.cc/WX6J-M2WU].

[101] Evelyn Douek, *Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" and Drowning Out Speech*, 94 AUSTL. L.J. 41, 45 (2020).

[102] *Id.* at 53.

[103] *See Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) (Austl.) [hereinafter *AVM Act*]; Online Safety Bill 2022-3 HC Bill [285] cl. 15 (UK) (requirement to preserve "content of democratic importance"); Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], Sept. 1, 2017, BUNDESGESETZBLATT, Teil I [BGBL. I] (Ger.) (requirement to block "manifestly unlawful" content within 24 hours); Regulation 2021/784 of the European Parliament and of the Council on Addressing the Dissemination of Terrorist Content Online, art. 3.3, 2021 O.J. (L 172) 79, 90 (obligation to remove terrorist content within one hour of receipt of removal order).

crisis were seen as unusually high are notable examples.[104] But platforms are increasingly applying this approach more broadly in high-risk contexts. Facebook has a set of "break glass" measures that it deploys "ahead of elections and during periods of heightened unrest," for example, that includes significantly reducing content that *likely* violates its policies.[105]

A natural reaction to hearing about these measures might be "why don't platforms take these risk-minimization measures all the time?" But the costs of too many false positives can also be substantial. Platforms censored posts from Palestinians raising awareness of their cause during violence in Gaza in summer of 2021, for example, in part due to more risk-averse content moderation measures in the region.[106] Evidence of war crimes in Syria was regularly scrubbed from the internet.[107] Regulators' notional sensitivity to the costs of over-moderating is evident in regulations that include carve-outs for categories such as "content of democratic importance" from takedown mandates.[108] But it's impossible to increase true positives at scale without also increasing false positives.

Technical and resource capacity will also be central to error choice. These factors will differ from platform to platform. This was discussed by Cloudflare when it made its child sexual abuse material detection technology available to all its clients. In a blog post, it discussed the tradeoffs in deciding the accuracy threshold at which to set the matching filter.[109] At any threshold there will be false negatives and false positives — in setting the threshold, platforms choose which to prefer. This might appear to be an easy choice: CSAM is so heinous that platforms should err on the side of false positives. But too loose and the number of hits will mean true positives will be lost in a sea of

[104] Evelyn Douek, *COVID-19 and Social Media Content Moderation*, LAWFARE (Mar. 25, 2020, 1:10 PM), https://www.lawfareblog.com/covid-19-and-social-media-content-moderation [https://perma.cc/8GRL-QR9Z].

[105] Miranda Sissons & Nicole Isaac, *Our Approach to Maintaining a Safe Online Environment in Countries at Risk*, FACEBOOK NEWSROOM (Oct. 23, 2021), https://about.fb.com/news/2021/10/approach-to-countries-at-risk [https://perma.cc/AMF2-X5PK]; Monika Bickert, *Preparing for a Verdict in the Trial of Derek Chauvin*, FACEBOOK NEWSROOM (Apr. 19, 2021), https://about.fb.com/news/2021/04/preparing-for-a-verdict-in-the-trial-of-derek-chauvin [https://perma.cc/YRA9-RH6Q].

[106] Bani Sapra, *Facebook's Algorithms Silenced Palestinian Voices. Can Its Biases Ever Be Fixed?*, WIRED (July 27, 2021), https://wired.me/business/big-tech/facebook-content-moderation-palestine [https://perma.cc/P5FZ-FWFW].

[107] Kate O'Flaherty, *YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks*, WIRED UK (June 26, 2018, 7:00 AM), https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video [https://perma.cc/G883-WWWF].

[108] Online Safety Bill 2022, cl. 15 (UK).

[109] 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/XAC4-VJE8].

errors.[110] Too tight and CSAM would escape the filter. Cloudflare concluded that the threshold should differ based on different platforms' size, capacity, and risk profile, and so let platforms choose their own threshold.[111] Even in the context of this simpler form of automated moderation (a matching technology for perhaps the most readily identifiable category of banned content), the level of risk depends on the broader content moderation system in which the tool is embedded.

These tradeoffs are even harder in other contexts where moderation technology is less accurate. The standard picture is filled with portraits of platform policymakers wrestling with how to balance the value of speech against other interests to arrive at a theoretically optimal rule, again mirroring offline constitutional judicial adjudication. But this idealistic picture is incomplete. Because of the practical challenges of moderating content at online speed and scale, whether a rule is realistically and technologically capable of being enforced is central to the rulemaking process. The level of nuance a hate speech–detection algorithm is capable of detecting, for example, can determine what substantive rules are possible. What if AI can't tell the difference between an image of a breast intended to raise awareness of breast cancer and one that is adult content? Principles are important, but every rule is tested for enforceability before it is promulgated.[112]

Ensuring rules can be enforced is a realistic approach to policymaking.[113] But there is something uncomfortable about such a functional approach to making speech rules ("We'd ban white nationalism, but the AI sucks at distinguishing it from certain forms of patriotism!"). Should rules be expressive and aspirational,[114] especially as rules themselves can set and shape social norms,[115] or should they reflect the bounds of what is possible in a particular moment, even if compromised? Neither principle nor practicality alone can answer what the "right" policy might

[110] *See, e.g.*, Kashmir Hill, *A Dad Took Photos of His Naked Toddler for the Doctor. Google Flagged Him as a Criminal.*, N.Y. TIMES (Aug. 25, 2022), https://www.nytimes.com/2022/08/21/technology/google-surveillance-toddler-photo.html [https://perma.cc/BZL7-XKED].

[111] Paine & Graham-Cumming, *supra* note 109.

[112] Bickert, *supra* note 36, at 257 ("Each component — the content reviewers, the automated systems, the policy writers, the engineers — depends on the other components for input and refinement."); *YouTube Community Guidelines*, *Developing Community Guidelines*, *How YouTube Develops Policies*, YOUTUBE (May 10, 2021), https://www.youtube.com/howyoutubeworks/policies/community-guidelines/#developing-community-guidelines [https://perma.cc/LE8A-RSSR] (describing, at 2:44 in the embedded video, YouTube's "four stage approval process" for policy development that depends on enforceability).

[113] Of course, in reality the gap between platform policies and actual enforcement leaves much to be desired.

[114] *See generally* Cass R. Sunstein, *On the Expressive Function of Law*, 144 U. PA. L. REV. 2021 (1996).

[115] *See generally* Lawrence Lessig, *The Regulation of Social Meaning*, 62 U. CHI. L. REV. 943 (1995).

be, and answers will be contingent on technological capacity. This pragmatism is hidden in the standard picture that is based around the formulation of ideal rules *before* their enforcement.

In sum, error choice is baked in at the moment of ex ante system design and depends on a number of factors including the importance of speed, an assessment of the level of risk in a particular context, and the level of technological capacity for moderating a certain kind of content. Each of these factors is not static and will require constant reevaluation.

2. Gatekeeping Versus User Control. — The standard picture depicts a binary of a platform as Governor and users as the governed. This reflects the moment in which the first wave of content moderation scholarship was written — the growth of major platforms was indeed a story of increasingly centralized platform control. This centralization in a handful of gatekeepers accelerated over the past half decade as lawmakers and the public demanded platforms to more actively exercise their power to police their services.[116] But there is now a countertrend. Centralization can leave users feeling disempowered and a lack of competition has constrained the ability of users to express dissatisfaction through exit, raising concerns about a handful of platforms' outsized power. As a result, there have been calls to return to the Internet's original promise and re-decentralize the web by empowering users to control their own experience.[117] This can happen through increased competition, new decentralized platforms, tools to bring decentralization to old platforms,[118] or, increasingly, large platforms giving users more control over content moderation.[119]

Community-based moderation reflects an ideal similar to the federalism principle of subsidiarity, allowing for experimentation and for online communities to exercise a more responsive form of governance. But it creates a new tradeoff: When users get content moderation

[116] *See* Jonathan Zittrain, Three Eras of Digital Governance 7 n.13, 8–9 (Sept. 23, 2019) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.3458435 [https://perma.cc/3XCY-3RMT]; Douek, *supra* note 77, at 776–84.

[117] *See* Liat Clark, *Tim Berners-Lee: We Need to Re-Decentralise the Web*, WIRED UK (Feb. 6, 2014, 1:38 PM), https://www.wired.co.uk/article/tim-berners-lee-reclaim-the-web [https://perma.cc/BY76-WNTJ]; Adi Robertson, *Twitter's Decentralized Social Network Project Takes a Baby Step Forward*, THE VERGE (Jan. 21, 2021, 3:34 PM), https://www.theverge.com/2021/1/21/22242718/twitter-bluesky-decentralized-social-media-team-project-update [https://perma.cc/4THD-QESS]; MIKE MASNICK, KNIGHT FIRST AMEND. INST., PROTOCOLS, NOT PLATFORMS: A TECHNOLOGICAL APPROACH TO FREE SPEECH 7 (2019), https://s3.amazonaws.com/kfai-documents/documents/e3288c9457/MasnickPublish.pdf [https://perma.cc/7M3S-MVAC]; Keller, *supra* note 13, at 26–27.

[118] MASNICK, *supra* note 117, at 14–26.

[119] *See supra* section II.A.4, pp. 545–48.

"wrong," should platforms intervene again?[120] Should the redistribution of power be contingent on obedience to certain substantive norms and, if so, who decides which ones? By encouraging such interventions, many lawmakers bemoan platform power while also asking platforms to exercise more of it and become more heavy-handed gatekeepers.

3. Procedural Justice and Legitimacy Tradeoffs at Scale. — The "techlash" against content moderation practices is driven by not simply substantive disagreement with platforms' decisions, but also the legitimacy deficits created by the lack of any public accountability or requirements of rationality for such decisions.[121] Of course, platforms are private actors and profit-driven companies. Under U.S. law, users have generally no claims of rights against them.[122] But clearly their decisions *do* affect public speech and democratic discourse on a massive scale. Thus, even without formal legal obligations, there are social expectations that these private decisionmakers should respect rule-of-law values.[123] Informally, then, content moderation systems are expected to observe some of the foundational values of public law, including procedural regularity and non-arbitrariness.[124]

Platforms entertain these expectations to a degree. So far, almost every transparency measure platforms have adopted has been voluntary, in response to regulatory and public pressure to explain themselves. Mark Zuckerberg has stated that "Facebook should not make so many important decisions about free expression and safety on our own."[125] Jack Dorsey acknowledged that Twitter making consequential decisions "with zero context as to why" is "unacceptable."[126] TikTok states that it releases transparency reports to "foster candid dialogue essential to

[120] *See, e.g.*, u/spez, *R/Announcements: Debate, Dissent, and Protest on Reddit*, REDDIT (Aug. 25, 2021), https://www.reddit.com/r/announcements/comments/pbmy5y/debate_dissent_and_protest_on_reddit [https://perma.cc/KB95-J85R]; *see also* DAPHNE KELLER, KNIGHT FIRST AMEND. INST., AMPLIFICATION AND ITS DISCONTENTS 35 (2021), https://s3.amazonaws.com/kfai-documents/documents/aa473e4dad/8.12.2021_-Keller-New-Layout.pdf [https://perma.cc/7S3F-Z96L]; Douek, *supra* note 59.

[121] Badiei et al., *supra* note 22, at 39 ("Establishing a platform's legitimacy might be harder than it is for other authorities that have been approved by their community members and are appointed through a democratic process.").

[122] Keller, *supra* note 13, at 11.

[123] Cass R. Sunstein & Adrian Vermeule, *The Morality of Administrative Law*, 131 HARV. L. REV. 1924, 1964 (2018) (describing how in paradigmatic cases of adjudication, decisionmakers have a set of intuitions that rule-of-law values should apply, even when there is no formal legal source of such requirements); Douek, *supra* note 77, at 820.

[124] *See* Giulio Napolitano, *The Rule of Law*, *in* THE OXFORD HANDBOOK OF COMPARATIVE ADMINISTRATIVE LAW 421, 426 (Peter Cane et al. eds., 2020) ("'Securing the Rule of Law' is now considered the 'core of administrative law.'" (quoting Richard B. Stewart, *Administrative Law in the Twenty-First Century*, 78 N.Y.U. L. REV. 437, 438 (2003))); *cf.* Jody Freeman & Sharon Jacobs, *Structural Deregulation*, 135 HARV. L. REV. 585, 634–35 (2021).

[125] Zuckerberg, *supra* note 80.

[126] *See* Jack Dorsey (@jack), TWITTER (Oct. 14, 2020, 7:55 PM), https://twitter.com/jack/status/1316528193621327876 [https://perma.cc/GF6B-Q4F9].

earning and maintaining trust" and that it is "incumbent" upon the company to provide clear information about how it protects user rights.[127] Transparency reporting is now largely standard across the industry, despite not being legally mandated, demonstrating that platforms acknowledge they have *some* public-regarding obligations.

Other efforts to bolster legitimacy have focused on providing users a measure of procedural justice. A report by a Facebook-commissioned group of academics at Yale University, including leading procedural justice scholar Professor Tom Tyler, recommended that, to garner trust and legitimacy, Facebook should provide more detailed reasoning to users when their posts are found to violate the community standards, more effective appeal rights, and greater transparency about the content moderation process.[128] Twitter has described how it aims to implement principles of procedural justice so users feel they have been treated fairly.[129] Notice-and-appeal regimes are increasingly common across the social media industry.

But pursuing legitimacy through procedural justice can be in tension with pursuing accuracy and efficiency. As discussed further below, it may be that greater emphasis on procedural justice for individual users decreases overall system accuracy and will have a complicated relationship with overall fairness by privileging certain interests over others.[130] Enshrining extensive procedural rights in law could also have adverse impacts on competition, if larger platforms are the only companies with the resources to comply.

The institutional design question of which procedural rights to afford users in individual cases is an ex ante choice with significant ramifications at the systems, not only individual user, level.

4. Consistency Versus Contextual Decisionmaking. — The picture of a transmission belt bureaucracy simply following rules from the standard picture is complicated by platforms' global scale.[131] One assumption content moderation system designers made in the early platform era was that American-style free speech norms should be universalized, and could be imposed on countries with different understandings of freedom of expression.[132] Since then, platforms have faced

[127] Eric Ebenstein, *Our First Transparency Report*, TIKTOK NEWSROOM (Dec. 30, 2019), https://newsroom.tiktok.com/en-us/our-first-transparency-report [https://perma.cc/A7UN-3RY7].

[128] BRADFORD ET AL., *supra* note 22, at 1, 34–39.

[129] *See* Twitter Safety, *What Procedural Justice Taught Us About Fairness and Rule Enforcement* (Sept. 21, 2021), https://blog.twitter.com/content/blog-twitter/common-thread/en/topics/stories/2021/what-procedural-justice-taught-us-about-fairness.html [https://perma.cc/4RED-RJX6].

[130] *See infra* section III.B.4, pp. 577–84.

[131] *See* CAPLAN, *supra* note 30, at 26.

[132] *See* Klonick, *supra* note 4, at 1621; Douek, *supra* note 77, at 771.

constant demands to acknowledge the unique concerns of varying contexts.[133] This creates a tradeoff between being attentive to local context and the rule-of-law value of consistency — the "right" answer in a paradigm case may in fact vary from country to country and case to case. In a sense, this is just another version of the tradeoff of the administrability and certainty of rules against the optimization of values-driven standards,[134] but at a global scale. Geopolitics complicates the picture further as platforms are put in the position of judging which states' local laws to obey and which they should resist in the name of the human rights of their users. The standard picture has little to say about this tension between the values of consistency and context sensitivity.

There are also tradeoffs regarding consistency of rules over time. The standard picture is a snapshot in time and suggests a certain level of stability in the system, but nothing about content moderation stands still.[135] The meaning of a hashtag or meme can change overnight, as when gay men responded to President Trump referring to the Proud Boys extremist group during a televised debate by reclaiming the hashtag #ProudBoys to make it an expression of gay love.[136] The tactics adversaries use to manipulate platforms are constantly changing in a game of cat-and-mouse, as when users add the words "breast cancer" to adult content to bypass automated tools trained to take down the latter but not the former.[137] The *very platforms themselves* are always changing — the affordances a platform offers for users to create, react to, or share content are in constant flux.[138] Frequent changes in rules make it

[133] *See, e.g.*, SIVA VAIDHYANATHAN, ANTISOCIAL MEDIA: HOW FACEBOOK DISCONNECTS US AND UNDERMINES DEMOCRACY 29 (2018) ("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 2–3 (Jan. 23, 2018) (unpublished manuscript), https://dx.doi.org/10.2139/ssrn.3108238 [https://perma.cc/PZ29-WSS5]; KAYE, *supra* note 22, at 117.

[134] *See* Frederick Schauer, *The Convergence of Rules and Standards*, 2003 N.Z. L. REV. 303, 309; *cf.* Kathleen M. Sullivan, *Post-Liberal Judging: The Roles of Categorization and Balancing*, 63 U. COLO. L. REV. 293, 306 (1992). *See generally* Duncan Kennedy, *Form and Substance in Private Law Adjudication*, 89 HARV. L. REV. 1685 (1976); Louis Kaplow, *Balancing Versus Structured Decision Procedures: Antitrust, Title VII Disparate Impact, and Constitutional Law Strict Scrutiny*, 167 U. PA. L. REV. 1375 (2019).

[135] James Grimmelmann, *The Platform Is the Message*, 2 GEO. L. TECH. REV. 217, 224–25 (2018).

[136] Lauren Strapagiel, *The "Proud Boys" Hashtag Has Been Taken Over by Gay Men*, BUZZFEED NEWS (Oct. 4, 2020, 4:49 PM), https://www.buzzfeednews.com/article/laurenstrapagiel/proud-boys-hashtag-reclaimed-gay-love-photos [https://perma.cc/Z6QR-RKE9].

[137] *See, e.g.*, NATHANIEL GLEICHER ET AL., FACEBOOK, THREAT REPORT: THE STATE OF INFLUENCE OPERATIONS 2017–2020, at 3 (2021), https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf [https://perma.cc/TR85-54L2] (describing how efforts to combat influence operations have "pressed threat actors to shift their tactics"); Nick Clegg, *Facebook's Response to the Oversight Board's First Set of Recommendations*, FACEBOOK NEWSROOM (Feb. 25, 2021), https://about.fb.com/news/2021/02/facebook-response-to-the-oversight-boards-first-set-of-recommendations [https://perma.cc/572S-QDL9].

[138] *See supra* section II.A.4, pp. 545–48.

hard for users to know what the rules are and create a specter of arbitrariness or irrationality.[139] To achieve stability, the standard picture suggests platforms should develop a body of precedent.[140] But what constitutes valuable precedent or sufficient notice to users in an ever-changing environment?[141] Platforms and regulators need to balance the desire for consistency and predictability of rules against the need to respond to changing circumstances and rapidly emerging threats and trends.

C. The Standard Picture's Mistaken Assumptions

Because content moderation involves speech and because so much of content moderation practice and academic discourse has been dominated by lawyers both within and outside platforms, such discourse is pervaded by First Amendment analogies.[142] Such analogies have been persistent, even as content moderation bureaucracies have outgrown them. These analogies, and the standard picture that invokes them, rely on mistaken assumptions about the necessary nature of speech governance. Such discourse assumes: (1) speech interests are special and especially resistant to systemic governance; (2) judicial-style ex post interventions must be individualistic; and (3) perfectibility is a necessary and desirable goal of speech regulation.

1. Speech Is Not So Special. — I am not the first to note the growing mismatch between constitutional-rights analogies and content moderation governance. Professor Jack Balkin presciently observed over a decade ago that "[p]rotecting free speech values in the digital age will be less and less a problem of constitutional law . . . and more and more a problem of technology and administrative regulation."[143] He later described how "new-school speech regulation" by platforms has become "in essence, a system of administrative law . . . . But the administrative agency in this case is a private company."[144] Similarly, Professor Kyle Langvardt hypothesized that the sheer scale of content moderation meant "a new 'management side' of free speech may come to eclipse the

[139] *Cf.* LON L. FULLER, THE MORALITY OF LAW 39 (rev. ed. 1969); Sunstein & Vermeule, *supra* note 123, at 1947–48.

[140] *See* Klonick, *supra* note 4, at 1645–46 (describing the importance of analogical reasoning); Kaye, *supra* note 99, at 19 (suggesting platforms develop a body of "case law").

[141] A Florida statute would have prohibited platforms from changing their rules more than once every thirty days and required them to notify each user about any changes. Fla. Stat. Ann. § 501.2041(2)(a), (c) (West 2022). This could thwart efforts to respond to emerging threats.

[142] *See* Klonick, *supra* note 4, at 1621; Ammori, *supra* note 22, at 2262.

[143] Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 427, 441 (2009).

[144] Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2028–29 (2018).

more familiar rights-based model in significance."[145] Others have noted that, as in many administrative agencies, platforms combine rulemaking, enforcement, and adjudication in the same hands.[146] Scholars have called on platforms to "turn to the principles and values of administrative law as a way of enhancing their own legitimacy" and make their systems of governance more accountable to the public.[147] Professor Rory Van Loo has explored "federal rules of platform procedure" with administrative oversight to regularize platform decisionmaking.[148]

But despite these earlier calls to look beyond individual cases to the systems and processes of content moderation, most discourse about it remains stuck in the register of individual ex post constitutional-rights litigation. Perhaps this should be unsurprising: content moderation decisions are decisions about *speech* after all. And speech, of course, is special.[149]

Particularly in the United States, it feels almost sacrilegious to suggest that speech should be administered systemically rather than treated as a sacred individual right deserving of the highest protection. There is perhaps no more emblematic and carefully guarded constitutional right than freedom of speech. Its *Firstness* is emphasized (even though it wasn't actually first in the original draft of the Bill of Rights),[150] and "[l]ike the sun, the First Amendment's size and brightness tends to blot out all else."[151] If constitutional law is generally resistant to incorporating ideas of systemic administration,[152] nowhere is this truer than in the domain of speech rights. As Professor Genevieve Lakier writes, "[t]he Free Speech Clause of the First Amendment has for decades now served as one of the most powerful mechanisms of individual rights protection in the Federal Constitution."[153] This is truer now more than ever, as the

[145] Langvardt, *supra* note 22, at 292 (footnote omitted) (quoting Jerry Mashaw, *The Management Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy, Fairness and Timeliness in the Adjudication of Social Welfare Claims*, 59 CORNELL L. REV. 772 (1974)).

[146] MACKINNON, *supra* note 22, at 154.

[147] Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 71 (2019); *see also* Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1062–69 (2018).

[148] Rory Van Loo, *Federal Rules of Platform Procedure*, 88 U. CHI. L. REV. 829, 895 (2021).

[149] Frederick Schauer, *Must Speech Be Special?*, 78 NW. U. L. REV. 1284, 1306 (1983).

[150] Albie Sachs, *Reflections on the Firstness of the First Amendment in the United States*, *in* THE FREE SPEECH CENTURY 179 (Lee C. Bollinger & Geoffrey R. Stone eds., 2018).

[151] Genevieve Lakier, *The Non–First Amendment Law of Freedom of Speech*, 134 HARV. L. REV. 2299, 2300–01 (2021).

[152] Gillian E. Metzger, *The Constitutional Duty to Supervise*, 124 YALE L.J. 1836, 1859 (2015).

[153] Lakier, *supra* note 151, at 2300–01.

rise of the Lochnerization of the First Amendment[154] has led to the Supreme Court striking down legislation across ever more domains.[155]

This elevation of speech rights is as much a cultural phenomenon as a legal one.[156] Content moderation debates are demonstrative. On the current state of the law, there is not even a colorable First Amendment claim against platforms for restricting users' speech.[157] Yet cries of "First Amendment!" or "Free Speech!" abound when they do. Against this cultural background, the hold of individual content moderation decisions on the collective imagination is strong and "one failure can incur enough public outrage to overshadow a million quiet successes."[158]

The example of eBay, which has largely escaped the techlash, illustrates how speech decisions stand apart. eBay handles over sixty million disputes annually, over ninety percent of them resolved by fully automated processes.[159] The platform's transparency reporting is perfunctory.[160] The stakes of a commercial transaction are arguably often higher than whether a post can remain on a single website. But as commercial decisions, eBay's governance is not thought of as "content moderation" or as implicating sacrosanct speech rights. eBay's biggest exposure to the techlash came only when the company chose to delist six Dr. Seuss books for including offensive imagery — that is, when it got into content moderation.[161] In fact, many canonical content moderation controversies are about commercial interests,[162] but they get

[154] Robert Post & Amanda Shanor, *Adam Smith's First Amendment*, 128 HARV. L. REV. F. 165, 165–67 (2015).

[155] *See, e.g.*, Reed v. Town of Gilbert, 135 S. Ct. 2218, 2237–38 (2015) (Kagan, J., concurring in the judgment).

[156] *See* 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/9D82-9KXL]; Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765, 1787–800 (2004).

[157] *See, e.g.*, Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019); Prager Univ. v. Google LLC, 951 F.3d 991, 995 (9th Cir. 2020). *But see* Biden v. Knight First Amend. Inst. at Columbia Univ., 141 S. Ct. 1220, 1221 (2021) (Thomas, J., concurring).

[158] GILLESPIE, *supra* note 14, at 9.

[159] Van Loo, *supra* note 4, at 559–60, 567; ETHAN KATSH & ORNA RABINOVICH-EINY, DIGITAL JUSTICE 34–35 (2017).

[160] *See* EBAY, 2021 GLOBAL TRANSPARENCY REPORT (2022), https://www.ebaymainstreet.com/sites/default/files/2021-05/2020-eBay-Global-Transparency-Report.pdf [https://perma.cc/3866-4FU6] (a generously spaced sixteen-page report of high-level data).

[161] Jeffrey A. Trachtenberg, *Dr. Seuss Books Deemed Offensive Will Be Delisted from eBay*, WALL ST. J. (Mar. 4, 2021, 6:11 PM), https://www.wsj.com/articles/dr-seuss-books-deemed-offensive-will-be-delisted-from-ebay-11614884201 [https://perma.cc/WBB8-M7CQ].

[162] *See, e.g.*, JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? 1–8 (2006) (discussing the famous controversy about Yahoo!'s sale of Nazi memorabilia in France); Richard Lawler, *OnlyFans Says Never Mind, It Actually Won't Ban Porn on October 1st*, THE VERGE (Aug.

framed as "speech" cases, making the "censored" party's grievance seem weightier. In a sense, every content moderation decision is commercial: private platforms are *profit-driven entities* that moderate because it is *in their business interests*.[163] But . . . speech!

Making progress on content moderation regulation requires getting past this speech squeamishness. There's no way to adequately reckon with the tradeoffs content moderation demands otherwise. Individual grievance obscures the need to talk about systems and compromises. Professor Rachel Barkow describes the same phenomenon in criminal law — sensational cases and headlines drive policy responses, preventing "rational discussion about whether, on balance, a particular program is a good idea. Instead, we get stories."[164] Barkow argues that these stories obscure policy tradeoffs and harness visceral reactions that skew risk tolerance.[165] Similarly, in content moderation, the idea of prioritizing the overall functioning of the system over individual rights is dissonant with the story American society tells itself about its free speech culture.

But fundamental rights are often administered systemically. Due process rights are the most obvious example. As Professor Richard Fallon observes, while "we characteristically think of constitutional rights in individualistic terms, due process doctrine has developed a strikingly managerial aspect."[166] Courts have also turned to managerial remedies where the interaction between different rights and interests is complicated.[167] Public law litigation, with its structural remedies for rights infringements, for example, has arisen in response to exactly the same concerns I raise in the next Part: individual relief is inefficient, distributively arbitrary, and unresponsive to broader, systemic interests in any claim.[168] In these circumstances, "it may be more efficient and more just to intervene preventatively, than to compensate post hoc."[169]

Technological change has been the impetus for many calls to move beyond an individualistic paradigm in the context of many rights.

25, 2021, 8:45 AM), https://www.theverge.com/2021/8/25/22640988/onlyfans-no-ban-porn-sexually-explicit-content-creators [https://perma.cc/2GGN-RX3V] (discussing OnlyFans's ban of adult content because of pressure from banking partners).

[163] GILLESPIE, *supra* note 14, at 11–12.

[164] Rachel E. Barkow, *Criminal Law as Regulation*, 8 N.Y.U. J.L. & LIBERTY 316, 319 (2014).

[165] *Id.* at 319–20.

[166] Richard H. Fallon, Jr., *Some Confusions About Due Process, Judicial Review, and Constitutional Remedies*, 93 COLUM. L. REV. 309, 311 (1993).

[167] *See, e.g.*, Martha Minow, *Judge for the Situation: Judge Jack Weinstein, Creator of Temporary Administrative Agencies*, 97 COLUM. L. REV. 2010, 2012 (1997); Charles F. Sabel & William H. Simon, *Destabilization Rights: How Public Law Litigation Succeeds*, 117 HARV. L. REV. 1016, 1020 (2004).

[168] Kathleen G. Noonan et al., *Reforming Institutions: The Judicial Function in Bankruptcy and Public Law Litigation*, 94 IND. L.J. 545, 557 (2019).

[169] *Id.*

Professor Daphna Renan has argued that the traditional individualistic framework of Fourth Amendment law is out of step with the increasingly programmatic nature of mass surveillance.[170] Doctrine that focuses on one-off encounters "fails to address — indeed, it cannot even see — . . . cumulative implications."[171] Relevantly, Renan argues that a key driver of this growing gap between Fourth Amendment doctrine and how modern surveillance practices work is the fact that technology has allowed for unprecedented aggregation and tractability of information.[172] This exact development is what requires rethinking speech governance too: the arrival of the internet and the development of automated moderation has made more speech possible, tractable, and regulable than ever before.[173] In the Fourth Amendment context, Renan argues this requires shifting from assessing particular cases to assessing structures and processes that protect the relevant interests in the aggregate and over time.[174] The argument of this Article is that this is true for content moderation too.

Analogous arguments have been made about the protection of privacy from private actors. Cohen has argued that privacy laws and regulatory proposals are similarly dominated by "a governance paradigm that is deeply embedded in the U.S. legal tradition and that relies on individual assertion of rights to achieve social goals" but does not account for "problems of design, networked flow, and scale."[175] Professor Margot Kaminski has likewise argued in this context that algorithmic decisionmaking requires systemic regulation and collaborative governance between public and private actors.[176]

This is a theme in the literature about governance of algorithmic decisionmaking generally.[177] Professor Aziz Huq observes that constitutional rights are largely enforced through discrete, individual legal actions, but the increased use of machine learning to make decisions affecting people's rights and interests means that normative implications

[170] Daphna Renan, *The Fourth Amendment as Administrative Governance*, 68 STAN. L. REV. 1039, 1041–42 (2016).

[171] *Id.* at 1059.

[172] *Id.* at 1056–57.

[173] Wu, *supra* note 34, at 2021.

[174] Renan, *supra* note 170, at 1081.

[175] JULIE E. COHEN, KNIGHT FIRST AMEND. INST., HOW (NOT) TO WRITE A PRIVACY LAW 3 (Mar. 23, 2021), https://s3.amazonaws.com/kfai-documents/documents/306f33954a/3.23.2021-Cohen.pdf [https://perma.cc/CE9P-CARY].

[176] Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. CAL. L. REV. 1529, 1533–34 (2019); *see also* Rory Van Loo, *The Missing Regulatory State: Monitoring Businesses in an Age of Surveillance*, 72 VAND. L. REV. 1563, 1565–66 (2019) (proposing monitoring of tech platforms and emphasizing the collaborative governance benefits of doing so).

[177] Kaminski, *supra* note 176, at 1540–41.

arise at the level of system design.[178] As Huq summarizes, "[w]ithout taking a systemic perspective that attends to the suite of human design decisions [associated with the use of AI], it will often not be feasible to identify how or why inaccuracies or systemic biases occur."[179]

The common thread here is that new technologies require new thinking about what it means to assert and protect rights. The scale and automated nature of decisionmaking have changed how decisions about rights are made by public and private entities alike. Scholars have argued that this requires rethinking governance in the context of due process,[180] antidiscrimination,[181] and privacy[182] rights. And yet speech governance proves especially resistant to this kind of analysis because, well, *speech*.

But there is nothing about speech decisions that makes an approach sensitive to the overall functioning of the system of speech governance distinctively or intrinsically inappropriate. Indeed, an oft-overlooked part of the First Amendment tradition takes precisely that approach. As Lakier describes it, "[f]or much of the twentieth century, the Court interpreted the guarantee of expressive equality in a manner that was sensitive to the economic, political, and social inequalities that inhibited or enhanced expression."[183] Perhaps most famously, Professor Alexander Meiklejohn argued that First Amendment doctrine should not protect individual autonomy but self-government, asserting that "[w]hat is essential is not that everyone shall speak, but that everything worth saying shall be said."[184] Structural considerations are a theme reflected in much recent scholarship anxious about the Lochnerization of the First Amendment. As Professors Jeremy Kessler and David Pozen ask, why should the focus be purely on speakers' interests "rather than ask[ing] which sorts of regulation would best serve the expressive environment as a whole?"[185] But this conception has not generally won in court.[186]

[178] Aziz Z. Huq, *Constitutional Rights in the Machine-Learning State*, 105 CORNELL L. REV. 1875, 1883–84 (2020).

[179] *Id.* at 1937.

[180] *See, e.g.*, Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1253 (2008); Huq, *supra* note 178, at 1905.

[181] *See, e.g.*, Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 CALIF. L. REV. 671, 675–76 (2016); Huq, *supra* note 178, at 1917–26.

[182] Kaminski, *supra* note 176; COHEN, *supra* note 175, at 2–3; Huq, *supra* note 178, at 1927–36.

[183] Genevieve Lakier, *Imagining an Antisubordinating First Amendment*, 118 COLUM. L. REV. 2117, 2119 (2018).

[184] ALEXANDER MEIKLEJOHN, FREE SPEECH AND ITS RELATION TO SELF-GOVERNMENT 25 (1948).

[185] Jeremy K. Kessler & David E. Pozen, *The Search for an Egalitarian First Amendment*, 118 COLUM. L. REV. 1953, 2001 (2018).

[186] 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, 1404–08 (2017).

The individualist conception of speech rights continues to dominate and trickles down to all discussion of speech interests, even in the context of private platforms.

But a less individualistic conception is not entirely theoretical or relegated to the history books. Lakier's account of what she calls non–First Amendment free speech law describes a body of American speech regulation that is more pluralistic and redistributive than is commonly acknowledged in the "constitutional register" of First Amendment law.[187] Local, state, and federal legislatures have long enacted laws "much more concerned with the threat that private economic power poses to expressive freedom,"[188] including common carrier obligations and other media regulation concerned with actively promoting free speech values.[189]

This more positive and pluralistic understanding of free speech fits with literature about the way technology has transformed the speech environment. Balkin presciently wrote about the cultural and participatory effects of digital technologies and the fact that "[i]ncreasingly, freedom of speech will depend on the design of the technological infrastructure that *supports the system* of free expression and secures widespread democratic participation."[190] Professor Tim Wu argues that the First Amendment and the assumptions it is based on are now obsolete.[191] Wu suggests that because technology has changed the world from one of informational to attentional scarcity, the First Amendment's solicitude of speech rights and government non-intervention makes it unprepared to deal with modern threats to public discourse.[192] And, as it turns out, the tractability of online speech due to technological advances has not made it any less unwieldy to govern — if anything, the opposite is the case.

In sum, the argument that the online environment requires a shift in our understanding of free speech to take a less individualistic view of speech rights is based in principle and pragmatism. And yet, content moderation debates remain largely stuck in an individualized constitutional register of the First Amendment, preoccupied with how to vindicate individual speech interests. This is misguided. Online speech interests can be effectively governed only by moving from focusing on individual to systemic concerns.

[187] Lakier, *supra* note 151, at 2308, 2371.

[188] *Id.* at 2304.

[189] *Id.* at 2330.

[190] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 6 (2004) (emphasis added).

[191] TIM WU, KNIGHT FIRST AMEND. INST., IS THE FIRST AMENDMENT OBSOLETE? 2–3 (2017), https://s3.amazonaws.com/kfai-documents/documents/5d8a0f848d/Is-the-First-Amendment-Obsolete-.pdf [https://perma.cc/WLQ5-ZVWD].

[192] *See generally id.*

2. *Ex Post Review Can Be Systemic.* — The standard picture not only understates the complexity of content moderation systems, but also relies on oversimplified notions of the possibilities of ex post judicial-style review. First Amendment analogies suggest an idealized image of decisionmaking that reflects the "deep-rooted historic tradition that everyone should have [their] own day in court."[193] This relies on a "classical view of the judicial role"[194] or administrative adjudication that typically imagines individualized hearings where claimants present their cause before a neutral decisionmaker.[195] Professor Lon Fuller famously declared that adjudication was not well suited for "polycentric" problems involving large groups of people where any decision would have widespread and unforeseeable consequences.[196]

But in fact, this idealized conception of the judicial role is increasingly unrepresentative. The legal system frequently requires courts and other decisionmakers to balance the interests involved rather than focusing purely on the parties before the court and maximizing individual procedural protections alone. From administrative law,[197] to public law litigation,[198] to mass torts,[199] to class actions and multidistrict litigation,[200] the law has evolved beyond individualistic adjudication in many areas to accommodate complex systemic concerns, in many cases requiring judges to act more as administrators than traditional judicial officers.[201] Nearly forty percent of cases in federal courts now proceed in some form of aggregated litigation.[202] Below I describe the many reasons this kind of decisionmaking would be more effective as a form of oversight of dynamic systems performing mass adjudication.[203] For now, the point is that the standard picture "assumes an implausibly rigid conception"[204] of ex post decisionmaking derived from the First Amendment context. Other, less individualized models are not only possible but also common and familiar.

[193] Samuel Issacharoff & John Fabian Witt, *The Inevitability of Aggregate Settlement: An Institutional Account of American Tort Law*, 57 VAND. L. REV. 1571, 1572 (2004) (quoting Ortiz v. Fibreboard Corp., 527 U.S. 815, 846 (1999)); *see also* Michael Sant'Ambrogio & Adam S. Zimmerman, *Inside the Agency Class Action*, 126 YALE L.J. 1634, 1645 (2017).

[194] Judith Resnik, *Managerial Judges*, 96 HARV. L. REV. 374, 376 (1982).

[195] Sant'Ambrogio & Zimmerman, *supra* note 193, at 1689.

[196] Lon L. Fuller, *The Forms and Limits of Adjudication*, 92 HARV. L. REV. 353, 395 (1978).

[197] *See* Walters v. Nat'l Ass'n of Radiation Survivors, 473 U.S. 305, 321 (1985) (quoting Mathews v. Eldridge, 424 U.S. 319, 344 (1976)); *see also* David Ames et al., *Due Process and Mass Adjudication: Crisis and Reform*, 72 STAN. L. REV. 1, 22 (2020).

[198] *See* Resnik, *supra* note 194, at 424–31; Noonan et al., *supra* note 168, at 566–68.

[199] *See* Issacharoff & Witt, *supra* note 193, at 1573.

[200] *See* David L. Noll, *MDL as Public Administration*, 118 MICH. L. REV. 403, 407 (2019).

[201] Abbe R. Gluck, *Unorthodox Civil Procedure: Modern Multidistrict Litigation's Place in the Textbook Understandings of Procedure*, 165 U. PA. L. REV. 1669, 1673 (2017).

[202] Sant'Ambrogio & Zimmerman, *supra* note 193, at 1640.

[203] *See infra* section III.B.3, pp. 572–77.

[204] Noonan et al., *supra* note 168, at 586.

To the extent such an approach provokes concerns about judicial legitimacy and institutional competence,[205] it should be remembered that so too does an outmoded and idealized notion of decision-making that ignores structural factors and underrepresented interests.[206] The sheer scale of content moderation should — and likely would, in any context apart from speech disputes — suggest a more aggregated approach.

3. The Pursuit of Perfectibility Is Misguided. — The natural corollary of the belief that individual redress is a necessary part of speech governance is the assumption that in every case of error, a functioning regulatory system should expend significant resources correcting it. Lawmakers, civil society, and academics may accept the inevitability of error in content moderation at scale, but emphasize that this means it should be possible to appeal any decision so that frontline errors can be remedied. The ethos behind these reform proposals is that content moderation systems are perfectible with enough resources and tiers of review. On this view, errors are an embarrassment, and any mistake that is not corrected through appeal is evidence of system failure.

This is an outgrowth of the individualistic rights framing that the First Amendment analogy suggests. First Amendment doctrine itself so anxiously seeks to avoid false positives that it protects against even the specter of chilling protected speech.[207] Courts will take months or years to resolve speech cases, finely parsing the details in an extraordinary effort to get them "right."[208] But even after all this effort, and no matter how much process and how many checks are piled on, there will always be disagreement about what outcomes in speech cases should be. Perfectibility in any system of speech regulation is illusory. The unfathomable scale of content moderation makes this all the more true. To believe otherwise is to force necessary tradeoffs into the shadows and suggest reforms that will be counterproductive.

III. MISGUIDED REFORM EFFORTS BASED ON THE STANDARD PICTURE

The incomplete and oversimplified nature of the standard picture is not just a theoretical problem — it has practical ramifications. This Part first illustrates how regulatory or self-regulatory content moderation reforms have adopted the individualized ex post review mode of

[205] *See, e.g.*, *id.* at 546; Redish & Karaba, *supra* note 42, at 110; Resnik, *supra* note 194, at 424.

[206] *See, e.g.*, Marc Galanter, *Why the "Haves" Come Out Ahead: Speculations on the Limits of Legal Change*, 9 LAW & SOC'Y REV. 95 (1974).

[207] Frederick Schauer, *Fear, Risk and the First Amendment: Unraveling the "Chilling Effect*,*"* 58 B.U. L. REV. 685, 705 (1978); JOHN HART ELY, DEMOCRACY AND DISTRUST: A THEORY OF JUDICIAL REVIEW 105 (1980).

[208] Mchangama, *supra* note 98.

error correction and accountability that follows from how content moderation is conceptualized in the standard picture. I then show how this approach will fail to bring accountability to the system as a whole.[209]

A. The Standard Picture's Influence

The standard picture's understanding of content moderation as a rough analogue of offline constitutional speech adjudication builds in an assumption that the best way to reform content moderation decisionmaking is by granting users a limited set of speech and procedural rights that they can wield against the platforms in individual rights–style adjudication.

This dynamic is the reason there has been so much scholarly effort in recent years to argue that content moderation must align with rule-of-law values that focus on individualized procedures, as in the "digital constitutionalism" literature.[210] In this tradition, a prominent strand of academic and civil society discourse asserts that platforms should provide every user with notice, an opportunity to be heard, an avenue to appeal to an independent reviewer, and reasons for any decision taken against them.[211] Much of this work was a product of a time when platforms offered no procedural protections at all and there was little understanding of how content moderation systems work.

But despite the fact that this work responds to an earlier era of content moderation, this discourse still provides the basis for much of lawmakers' understanding of content moderation. As Professor Eric Goldman catalogues, regulation around the world has codified the assumption that content moderation is almost singularly concerned with the binary decision to take down or leave up individual pieces of content.[212] It ignores the institutional diversity and ex ante design choices platforms make. An early wave of online speech regulation focused on incentivizing platforms to take down content that regulators deemed

[209] *See* Freeman, *supra* note 10, at 664–66 (describing the notion of "aggregate accountability" for a system as a whole).

[210] *See, e.g.*, Nicolas Suzor, *Digital Constitutionalism: Using the Rule of Law to Evaluate the Legitimacy of Governance by Platforms*, SOC. MEDIA + SOC'Y, July–Sept. 2018, at 2; Claudia Padovani & Mauro Santaniello, *Digital Constitutionalism: Fundamental Rights and Power Limitation in the Internet Eco-System*, 80 INT'L COMMC'N GAZETTE 295, 296 (2018); Dennis Redeker et al., *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*, 80 INT'L COMMC'N GAZETTE 302, 303 (2018); Edoardo Celeste, *Digital Constitutionalism: A New Systematic Theorisation*, 33 INT'L REV. L. COMPUTS. & TECH. 76, 76–77 (2019); Giovanni De Gregorio, *The Rise of Digital Constitutionalism in the European Union*, 19 INT'L J. CONST. L. 41, 41 (2021).

[211] *See* Suzor, *supra* note 210, at 7–8; *see, e.g.*, *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, SANTA CLARA PRINCIPLES, https://santaclaraprinciples.org [https://perma.cc/ZF3J-UFC7]; BRADFORD ET AL., *supra* note 22, at 34–39; Emma J Llansó, *No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem*, BIG DATA & SOC'Y, Jan.–June 2020, at 4; Klonick, *supra* note 23, at 2479.

[212] Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1, 16 (2021).

harmful and punishing individual failures to do so.[213] But after much criticism and adverse court decisions,[214] regulators have attempted to mitigate such concerns by also mandating platforms provide individual due process rights. Now regulators try to have their cake and eat it too. The trend is to demand takedowns in ever-shorter timeframes without acknowledging the costs this will have for accuracy, sensitivity to context, or the practicality of individualized due process.[215]

The European Union Digital Services Act[216] (DSA) (and the Digital Services Oversight and Safety Act[217] (DSOSA) in the United States, which strongly resembles the DSA[218]) illustrates where this argument leads. It includes requirements for extensive procedural protections in every case: platforms would need to provide reasons for any content removal,[219] a right of appeal open for six months in all cases,[220] a human in the loop for all appeals,[221] and a further right of appeal to a third-party arbitrator.[222] Similarly, in the United States, the proposed Platform Accountability and Consumer Transparency Act[223] requires individual notice, appeals, and reasons.[224] Laws in Texas and Florida require the same.[225] Courts are (unsurprisingly, given this resembles their own approach) sympathetic to this approach too — the German Federal Court of Justice has ruled that Facebook must provide every

[213] *See, e.g.*, sources cited *supra* note 103.

[214] *See, e.g.*, Magyar Tartalomszolgáltatók Egyesülete & Index.hu Zrt v. Hungary, App. No. 22947/13, ¶ 82 (Feb. 5, 2016), https://hudoc.echr.coe.int/fre?i=001-160314 [https://perma.cc/G3EH-6UA5] (holding that compelling platforms to find and remove every unlawful user comment "amounts to requiring excessive and impracticable forethought capable of undermining freedom of the right to impart information on the Internet"); Conseil constitutionnel [CC] [Constitutional Court] decision No. 2020-801 DC, June 18, 2020, Rec. 10–11 (Fr.) (a decision of the French Constitutional Court striking down a French law directing platforms to take down hate speech in very short timeframes); Shreya Singhal v. Union of India, (2013) 12 SCC 73 (India) (striking down an Indian law creating liability for platforms that did not take down certain speech).

[215] *See* sources cited *supra* note 103.

[216] *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC*, COM (2020) 825 final (Sept. 7, 2022) [hereinafter *DSA*]. I will return to the DSA below as an example of regulation that incorporates provisions aimed at the broader systems and processes of content moderation beyond individual cases. *See infra* section IV.B, pp. 593–603.

[217] Digital Services Oversight and Safety Act of 2022, H.R. 6796, 117th Cong. (2022).

[218] *See, e.g.*, *id.* § 7 (risk assessment and risk mitigation reporting).

[219] *DSA*, *supra* note 216, art. 17.

[220] *Id.* art. 20.

[221] *Id.* art. 20(6).

[222] *Id.* art. 21.

[223] Platform Accountability and Consumer Transparency Act, S. 4066, 116th Cong. (2020) [hereinafter PACT Act].

[224] *Id.* § 5(c)(2).

[225] An Act Relating to Censorship of or Certain Other Interference with Digital Expression, 2021 Tex. Gen. Laws (codified at TEX. BUS. & COM. CODE ANN. §§ 120.001–.151 (2021) and TEX. CIV. PRAC. & REM. CODE ANN. §§ 143A.001–.008 (2021)); Social Media Platforms, 2021 Fla. Laws 32 (codified at FLA. STAT. §§ 106.072, 287.137, 501.2041, 501.212 (2021)).

individual user notice if a post is deleted, reasons for that deletion, and an opportunity to reply.[226]

These models do not acknowledge all the diverse design choices outlined above.[227] In drafting these laws, lawmakers also fail to acknowledge the diversity of content moderation in practice: the same executives from a few large companies are hauled before lawmakers again and again to account for the content moderation industry as a whole.[228] Lawmakers are therefore considering regulation based on a small slice of the decisionmaking within companies that in turn make up a small slice of the content moderation industry.

This more limited understanding of the practice of content moderation suits these larger platforms just fine though — a narrow understanding of content moderation leads to narrow regulation. Larger platforms will also most easily bear the burden of complying with mandates to provide individual users with greater procedural protections. Obtaining regulations that reflect the current dominant operating models favors incumbents, a kind of "regulatory capture through design."[229]

Nothing exemplifies the way the standard picture leads to ex post judicial review–style solutions — and platforms' encouragement of this framing — better than the Facebook Oversight Board. A high-profile experiment in self-regulation, the Oversight Board sits at the apex of Facebook's content moderation system and is colloquially known as Facebook's "Supreme Court."[230] As Professor Kate Klonick's leading account argues, "[t]he analogy to a constitution that guarantees substantive and procedural rights through review by an independent judiciary

[226] *Der Bundesgerichtshof–Presse: Pressemitteilungen aus dem Jahr 2021–Bundesgerichtshof zu Ansprüchen Gegen die Anbieterin Eines Sozialen Netzwerks, Die Unter dem Vorwurf der "Hassrede" Beiträge Gelöscht und Konten Gesperrt Hat*, BUNDESGERICHTSHOF (July 29, 2021), https://www.bundesgerichtshof.de/SharedDocs/Pressemitteilungen/DE/2021/2021149.html [https://perma.cc/Y83X-UY6F].

[227] Mike Masnick, *The Internet Is Not Just Facebook, Google & Twitter: Creating a "Test Suite" for Your Great Idea to Regulate the Internet*, TECHDIRT (Mar. 18, 2021, 9:38 AM), https://www.techdirt.com/articles/20210317/23530146442/internet-is-not-just-facebook-google-twitter-creating-test-suite-your-great-idea-to-regulate-internet.shtml [https://perma.cc/BA5S-84ZQ].

[228] Jay Peters, *2020 Is Giving Us Another Chance to Watch Mark Zuckerberg and Sundar Pichai Get Grilled by Congress*, THE VERGE (Oct. 2, 2020, 7:46 PM), https://www.theverge.com/2020/10/2/21499502/facebook-google-twitter-ceos-testify-senate-commerce-committee-october [https://perma.cc/7TP9-LVB9].

[229] Andrew I. Gavil, Essay, *The FTC's Study and Advocacy Authority in Its Second Century: A Look Ahead*, 83 GEO. WASH. L. REV. 1902, 1912 (2015).

[230] *See, e.g.*, Shira Ovide, *Facebook Invokes Its "Supreme Court,"* N.Y. TIMES (Oct. 21, 2021), https://www.nytimes.com/2021/01/22/technology/facebook-oversight-board-trump.html [https://perma.cc/R46C-J3XF] (originally published Jan. 22, 2021). I myself echoed Zuckerberg's use of the analogy in my early writing on the topic. *See* Evelyn Douek, *Facebook's New "Supreme Court" Could Revolutionize Online Speech*, LAWFARE (Nov. 19, 2018, 3:09 PM), https://www.lawfareblog.com/facebooks-new-supreme-court-could-revolutionize-online-speech [https://perma.cc/2SX8-RJFZ].

is crucial for understanding" the Board.[231] Others have compared the body to an international human rights tribunal.[232] Civil society has advocated for a similar cross-industry proposal in the form of "Social Media Councils."[233] Another proposal suggests online "e-courts" to review takedown decisions.[234] The tenor of the Board's decisions are overwhelmingly the kind of ex post review one would expect from a judicial body. The Board's mandate is confined to decisions about individual posts, and does not extend to review of any of the decisions outlined in section II.A above.[235] The Board's opinions turn on a very careful parsing of individual posts — even different translations of posts[236] — highly specific to their unique factual context, under Facebook's community standards, values, and international human rights norms. The Board's procedural expectations of Facebook epitomize the individual rights paradigm — a focus on providing notice, reasons, and an individual appeal to a human in every case.[237]

In this way, the Board's decisions reflect a common theme of content moderation discourse.[238] To the extent that content moderation in practice is seen as departing from the standard picture, the solution is generally to call for greater adherence to it: *Make sure there's always an avenue of appeal. Give frontline decisionmakers more context so they can more judiciously consider the case before them. Afford users more procedural justice.* There is an assumption that platforms *could* apply their rules correctly — that there *is* a "correct" application of the rules — if only content moderation bureaucracies were adequately resourced.

B. *The Futility and Failures of Individualized Ex Post Review*

Content moderation cannot be made accountable post by post. The individualistic ex post approach to error correction that flows from the

[231] Klonick, *supra* note 23, at 2477.

[232] Laurence Helfer & Molly K. Land, *Is the Facebook Oversight Board an International Human Rights Tribunal?*, LAWFARE (May 13, 2021, 8:01 AM), https://www.lawfareblog.com/facebook-oversight-board-international-human-rights-tribunal [https://perma.cc/V8MC-QDDR].

[233] ARTICLE 19, *supra* note 22.

[234] TRANSATLANTIC HIGH LEVEL WORKING GRP. ON CONTENT MODERATION ONLINE & FREEDOM OF EXPRESSION, FREEDOM AND ACCOUNTABILITY: A TRANSATLANTIC FRAMEWORK FOR MODERATING SPEECH ONLINE 27–28 (2020).

[235] OVERSIGHT BD., OVERSIGHT BOARD BYLAWS art. I, § 3 (2022), https://www.oversightboard.com/sr/governance/bylaws [https://perma.cc/629G-C85W]; *id.* art. III, § 1.1.1.

[236] *See, e.g.*, *Case Decision 2020-002-FB-UA*, OVERSIGHT BD. (Jan. 28, 2021), https://oversightboard.com/decision/FB-I2T6526K [https://perma.cc/3BHT-7RGK].

[237] Evelyn Douek, *The Facebook Oversight Board's First Decisions: Ambitious, And Perhaps Impractical*, LAWFARE (Jan. 28, 2021, 11:23 AM), https://www.lawfareblog.com/facebook-oversight-boards-first-decisions-ambitious-and-perhaps-impractical [https://perma.cc/WY43-Q5ET].

[238] *See, e.g.*, ASH ET AL., *supra* note 22, at 11–13; BRADFORD ET AL., *supra* note 22, at 8–10; Suzor, *supra* note 210; *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211.

standard picture is misguided for four main reasons: (1) individual cases are poor vehicles for identifying or reviewing systemic errors and ex ante design choices; (2) to the extent that these errors are identifiable, ex post individual remedies will be ineffective in reforming broken systems; (3) transparency oriented around individual cases will produce limited insight into the overall dynamics of a content moderation system; and (4) a simple maximalist understanding of due process rights does not engage with the tradeoffs involved in the provision of additional individual procedural protections.

1. Individual Errors Are Poor Vehicles for Identifying Systemic Failures. — Case-by-case review is a poor model of content moderation oversight because such review will, first, fail to identify systemic failures and, second, skew risk tolerance by highlighting mistakes that may be the product of reasonable ex ante decisions at the systems level.

First, a post-by-post approach "cannot even see" aggregate harms.[239] A system's disparate impact cannot be identified by looking at a single decision, and this is especially true for automated moderation. As Huq observes, "individual cases of erroneous decisions provide[] limited evidence that a particular algorithmic classification system has deviated from due process norms [or creates] . . . an equality-related problem."[240]

The example of hate speech, which most major platforms now prohibit, is illustrative. There are persistent concerns that platforms disproportionately enforce hate speech rules against marginalized groups and minorities.[241] These concerns seem to be somewhat justified. After years of dismissing complaints of bias, Facebook announced that it would overhaul its hate speech detection algorithms to address the disparate impact of its "race-blind" systems.[242] Perspective API, a tool used by a wide variety of platforms including Reddit and Latin America's second-largest platform, Taringa!,[243] has been shown to disproportionately flag Black users' speech as hate speech.[244] Ex post individualistic

[239] *See* Renan, *supra* note 170, at 1059; *see also* COHEN, *supra* note 84, at 153; Huq, *supra* note 178, at 1937.

[240] Huq, *supra* note 178, at 1937.

[241] LAURA W. MURPHY ET AL., FACEBOOK'S CIVIL RIGHTS AUDIT — FINAL REPORT 57 (2020).

[242] Elizabeth Dwoskin et al., *Facebook to Start Policing Anti-Black Hate Speech More Aggressively than Anti-White Comments, Documents Show*, WASH. POST (Dec. 3, 2020, 8:00 AM), https://www.washingtonpost.com/technology/2020/12/03/facebook-hate-speech [https://perma.cc/7CJA-QTXE].

[243] *Case Studies*, PERSPECTIVE API, https://www.perspectiveapi.com/case-studies [https://perma.cc/YJY9-JL49]; Jigsaw, *How Latin America's Second Largest Social Platform Moderates More than 150K Comments a Month*, MEDIUM (Aug. 29, 2019), https://medium.com/jigsaw/how-latin-americas-second-largest-social-platform-moderates-more-than-150k-comments-a-month-df0d8a3ac242 [https://perma.cc/J5TJ-9W5Z].

[244] Katyanna Quach, *Oh Dear . . . AI Models Used to Flag Hate Speech Online Are, Er, Racist Against Black People*, THE REGISTER (Oct. 11, 2019, 11:00 AM), https://www.theregister.co.uk/2019/10/11/ai_black_people [https://perma.cc/2KMT-3JUC].

review will treat individual outcomes of such bias as errors to be corrected rather than potential evidence of the need for systemic reform.

Not all systemic failures are the result of broken AI. Errors may be downstream from other forms of problematic system design that will not be apparent within the four corners of any individual decision. An example is Facebook's choice to leave up an event page belonging to a militia group called Kenosha Guard, which had issued a "call to arms" before a protest in Kenosha that turned deadly.[245] This might be framed as an individual failure by Facebook to enforce its rules prohibiting incitement — an "operational mistake," as Zuckerberg called it.[246] From this perspective, the error is a regrettable but inevitable product of operating at scale, to be corrected ex post.

But Facebook's failure in Kenosha was a failure of system design. The Kenosha Guard's event was reported over 455 times, making up a staggering sixty-six percent of event reports that day.[247] Failing to properly review an item responsible for two-thirds of a day's reports is not a story of unmanageable scale — it is a systemic breakdown.

Conversely, an error *may* be evidence of poor system design, but a single error may not be proof of system failure at all. Because platform scale means even the most carefully designed system will make mistakes,[248] an error might be the consequence of a calculated and reasonable ex ante tradeoff between differing values.

This is Barkow's insight in the criminal law context. Reacting to a single story produces solutions that may have greater social costs in the long run.[249] Emotive stories of platforms wronging sympathetic users are legion. Platforms have interfered with volunteer efforts to make and distribute masks when enforcing a mask ad ban at the start of the pandemic;[250] removed history videos when trying to purge Holocaust

[245] Russell Brandom, *Facebook Chose Not to Act on Militia Complaints Before Kenosha Shooting*, THE VERGE (Aug. 26, 2020, 5:15 PM), https://www.theverge.com/2020/8/26/21403004/facebook-kenosha-militia-groups-shooting-blm-protest [https://perma.cc/4ANV-PGPX].

[246] Ryan Mac, *Facebook Employees Are Outraged at Mark Zuckerberg's Explanations of How It Handled the Kenosha Violence*, BUZZFEED NEWS (Aug. 28, 2020, 3:10 PM), https://www.buzzfeednews.com/article/ryanmac/facebook-employees-slam-zuckerberg-kenosha-militia-shooting [https://perma.cc/J5ME-DJBW].

[247] 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, 6:45 PM), https://www.buzzfeednews.com/article/ryanmac/kenosha-militia-facebook-reported-455-times-moderators [https://perma.cc/HRE6-5PAA].

[248] *Cf.* Andrew D. Selbst, *An Institutional View of Algorithmic Impact Assessments*, 35 HARV. J.L. & TECH. 117, 138 (2021).

[249] Barkow, *supra* note 164, at 319; *see also* OMRI BEN-SHAHAR & CARL E. SCHNEIDER, MORE THAN YOU WANTED TO KNOW: THE FAILURE OF MANDATED DISCLOSURE 144 (2014).

[250] 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 [https://perma.cc/R6BP-54V7].

denial;[251] deplatformed antiracist skinheads when removing racist skinheads;[252] deleted obviously satirical cartoons when enforcing rules against advocating violence and spreading misinformation;[253] and suspended an account for posting a cartoon of Captain America punching a Nazi because it depicted, well, a Nazi.[254] Platforms admit these are mistakes. But looking at these stories in isolation ignores that errors are often caused by competing demands, such as that platforms should enforce nuanced rules quickly at scale. The crucial question must be: are these mistakes a reasonable cost for enforcing rules quickly and correctly in many (many!) other cases? But this is not a question that gets asked in most ex post reviews of individual content moderation decisions. This is because, to echo Barkow's conclusion in another context, content moderation is seen as a collection of stories that "engender visceral reactions" rather than a project of mass administration.[255] This causes the public and lawmakers to "fail to appreciate that there are tradeoffs and downsides when they pursue their goals."[256]

To understand a system overall, the relevant questions are not *if* there are errors but "[w]hich types of errors are known and tolerated, how is risk distributed, and who has the institutional standing or technological power" to alter these choices?[257]

2. Case-by-Case Review Provides Inadequate Remedies. — Individualistic ex post review focuses on rectification for the individual concerned, generally to the exclusion of others who were or will be similarly situated. As Cohen has written in the related context of privacy, "[a]tomistic, post hoc assertions of individual control rights . . . cannot meaningfully discipline networked processes that operate at scale. Nor can they reshape earlier decisions about the design of algorithms and user interfaces."[258] But to prevent future failures, "[c]orrection of error may require reformulation of the organization's strategy, not just the

[251] Elizabeth Dwoskin, *How YouTube Erased History in Its Battle Against White Supremacy*, WASH. POST (June 13, 2019, 12:55 PM), https://www.washingtonpost.com/technology/2019/06/13/how-youtube-erased-history-its-battle-against-white-supremacy [https://perma.cc/UM4C-UZH9].

[252] Chloe Hadavas, *Why We Should Care that Facebook Accidentally Deplatformed Hundreds of Users*, SLATE (June 12, 2020, 2:12 PM), https://slate.com/technology/2020/06/facebook-anti-racist-skinheads.html [https://perma.cc/W7JM-YAD9].

[253] Mike Isaac, *For Political Cartoonists, The Irony Was that Facebook Didn't Recognize Irony*, N.Y. TIMES (June 10, 2021), https://www.nytimes.com/2021/03/19/technology/political-cartoonists-facebook-satire-irony.html [https://perma.cc/4LHY-BHUW].

[254] Blake Montgomery, *Twitter Suspends an Account for a Cartoon of Captain America Punching a Nazi*, DAILY BEAST (Sept. 11, 2019, 11:35 AM), https://www.thedailybeast.com/twitter-suspends-an-account-for-tweeting-a-cartoon-of-captain-america-punching-a-nazi [https://perma.cc/C6WX-J3PL].

[255] Barkow, *supra* note 164, at 320.

[256] *Id.* at 328.

[257] Ananny, *supra* note 97.

[258] COHEN, *supra* note 175, at 4.

correction of departures from it."[259] To return to the example of Facebook's mistake in Kenosha, in isolation the problem could be resolved by banning the group after the fact. But this does not address inadequate systems leading to the mistake in the first place.

Given that any content moderation appeals system will only ever review the tiniest fraction of a platform's decisions, an ex post approach to correcting individual mistakes is tinkering at the margins. Furthermore, an ex post focus inevitably evaluates problems at a specific point in time, when the decision was made.[260] But content curation by platforms is continuous and rules and systems are updated on an "ongoing basis."[261] Content moderation must respond to evolving trends, which can emerge and become widespread within hours, and companies often take time-limited exceptional measures (for example, reducing the amplification of certain types of content or expanding the kinds of content they will remove) in particularly volatile situations.[262] A case-by-case assessment does not evaluate the effectiveness of such measures, relying on episodic intervention rather than ongoing oversight.[263] A backward-looking approach to regulating such a volatile environment will suggest solutions that are dated before they are implemented.

3. Transparency Theater. — The apparent simplicity of the standard picture leads regulators to adopt straightforward transparency mandates that in practice provide little meaningful insight. Transparency is a means, not an end, and needs to be targeted to be effective.[264] In the context of content moderation, however, as Daphne Keller put it, everyone agrees "more transparency is better. But . . . almost no one has a

[259] Charles F. Sabel & William H. Simon, *The Management Side of Due Process in the Service-Based Welfare State*, *in* ADMINISTRATIVE LAW FROM THE INSIDE OUT: ESSAYS ON THEMES IN THE WORK OF JERRY L. MASHAW 63, 83 (Nicholas R. Parrillo ed., 2017).

[260] *See* William H. Simon, *The Organizational Premises of Administrative Law*, 78 LAW & CONTEMP. PROBS. 61, 69 (2015).

[261] Keller, *supra* note 13, at 14.

[262] *See, e.g.*, Jessica Guynn, *Facebook Deploys Emergency Measures to Curb Misinformation as Nation Awaits Election Results*, USA TODAY (Nov. 6, 2020, 12:06 PM), https://www.usatoday.com/story/tech/2020/11/05/facebook-election-misinformation-crackdown-emergency-measures-trump/6182001002 [https://perma.cc/5X2M-6JFS]; Kate Conger, *Twitter Will Turn Off Some Features to Fight Election Misinformation*, N.Y. TIMES (Oct. 9, 2020), https://www.nytimes.com/2020/10/09/technology/twitter-election-ban-features.html [https://perma.cc/B6E4-L7DZ]; Leslie Miller, *How YouTube Supports Elections*, YOUTUBE OFF. BLOG (Feb. 3, 2020), https://blog.youtube/news-and-events/how-youtube-supports-elections [https://perma.cc/79L6-7XT3].

[263] *Cf.* Mark Tushnet, *Introduction: The Pasts & Futures of the Administrative State*, DAEDALUS, Summer 2021, at 5, 6.

[264] David E. Pozen, *Seeing Transparency More Clearly*, 80 PUB. ADMIN. REV. 326, 327 (2020).

clear wish list."[265] The standard picture's focus on paradigm cases suggests transparency should focus on how such cases are resolved by frontline decisionmakers and on appeal. But platform disclosures in this vein show the risks of transparency theater that obscures rather than illuminates.[266] Platforms can drown observers in data while revealing little.[267]

The dominant form of transparency mandate requires platforms to report gross enforcement numbers. This is no doubt in part because such a mandate is easy for a regulator to write.[268] But aggregate enforcement numbers, without more, do not explain relevant denominators or the cause of various trends. For example, when a platform reports an increase in takedowns, it might be intuitive to assume this is because that platform is doing a better job of finding and removing violating content. But there could be many other reasons: there could be more content overall on the platform, or an increase in that *kind* of content; the platform might have lowered its confidence threshold for removing violating content; the platform might have broadened its definition of violating content; and so on. Aggregate figures also do not provide insight into other important factors, like whether there is consistency across different languages, populations, and regions.

This kind of transparency is also modeled on what the largest platforms *already report*. In the words of Mark Zuckerberg: "Facebook already publishes transparency reports on how effectively we're removing harmful content. I believe every major Internet service should do this quarterly."[269] Facebook is so keen on this proposal that it spent more than any other big tech company on lobbying in 2020 to support updated regulations.[270] Why? These reports reveal little but would be costly for every platform to produce. *Of course* Facebook would love the law to require other platforms to also do what Facebook already does.

[265] Daphne Keller, *Some Humility About Transparency*, STAN. L. SCH. CTR. FOR INTERNET & SOC'Y (Mar. 19, 2021, 3:09 AM), http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency [https://perma.cc/J2SR-3GQB].

[266] Nicolas P. Suzor et al., *What Do We Mean when We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 INT'L J. COMMC'N 1526, 1528–29 (2019).

[267] *See* Sun-ha Hong, *Why Transparency Won't Save Us*, CTR. FOR INT'L GOVERNANCE INNOVATION (Feb. 18, 2021), https://www.cigionline.org/articles/why-transparency-wont-save-us [https://perma.cc/5M4X-S8K8]; Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 979 (2018).

[268] For a helpful discussion of these challenges, see Keller, *supra* note 265.

[269] Mark Zuckerberg, Opinion, *The Internet Needs New Rules. Let's Start in These Four Areas.*, WASH. POST (Mar. 30, 2019, 3:00 PM), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html [https://perma.cc/HD23-5XCQ].

[270] Lauren Feiner, *Facebook Spent More on Lobbying than Any Other Big Tech Company in 2020*, CNBC (Jan. 22, 2021, 11:41 AM), https://www.cnbc.com/2021/01/22/facebook-spent-more-on-lobbying-than-any-other-big-tech-company-in-2020.html [https://perma.cc/M2U6-E9P5].

The flaws of such specific, static transparency mandates are illustrated by Germany's NetzDG,[271] the first piece of legislation in the world to require content moderation transparency reports.[272] Despite widespread criticism of the law,[273] even detractors were "typically positive about NetzDG's transparency requirements,"[274] which require platforms to produce half-yearly reports including the number of complaints received about unlawful content, the source of those complaints, the action taken in response, and the length of time it took the platform to act.[275]

These mandates have underdelivered. The goal of the NetzDG reports was to provide information "necessary in the interest of an effective impact assessment" of the law,[276] but the lack of specificity in the reporting obligations means it is "almost impossible to truly evaluate the impact of such regulation."[277] Measuring success purely on the number of takedowns "may simply encourage over-removal."[278] Without any quality assurance about decisions or standardization across different platforms, it is impossible for "third parties to evaluate the merits of platform[s'] decisions" as to whether content did, in fact, violate German law or if NetzDG's requirements have had any effect on hate speech on the platform.[279]

The NetzDG reporting obligations also illustrate the standard picture's blindness to content moderation heterogeneity. The law's transparency mandate covers only one type of platform intervention: takedowns based on a discrete complaint. The Act ignores all the other types of content moderation described in section II.B or the possibility of new forms of intervention in the future.

Focusing on the wrong metrics is worse than uninformative: it creates perverse incentives. This format incentivizes platforms to simply report ever larger numbers, which is what regulators come to demand

[271] NetzDG, *supra* note 103.

[272] Robert Gorwa & Timothy Garton Ash, *Democratic Transparency in the Platform Society*, *in* SOCIAL MEDIA AND DEMOCRACY 286, 299–300 (Nathaniel Persily & Joshua A. Tucker eds., 2020).

[273] Evelyn Douek, *Germany's Bold Gambit to Prevent Online Hate Crimes and Fake News Takes Effect*, LAWFARE (Oct. 31, 2017, 11:30 AM), https://www.lawfareblog.com/germanys-bold-gambit-prevent-online-hate-crimes-and-fake-news-takes-effect [https://perma.cc/HKU2-UEVQ]; Amélie Heldt, *Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports*, INTERNET POL'Y REV., June 12, 2019, at 1, 2 (2019); Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law* 11 (Transatlantic Working Grp., Working Paper, 2019).

[274] Tworek & Leerssen, *supra* note 273, at 3.

[275] NetzDG, *supra* note 103, § 2.

[276] Heldt, *supra* note 273, at 7 (quoting Gesetzentwurf [Draft Bill], Deutscher Bundestag: Drucksachen [BT] 18/12356 (Ger.), at 20, https://dserver.bundestag.de/btd/18/123/1812356.pdf [https://perma.cc/Y85N-RCBM]).

[277] *Id.* at 14.

[278] Tworek & Leerssen, *supra* note 273, at 7.

[279] *Id.* at 8.

and what platforms boast about achieving. This illustrates the universal law that, left to their own devices, any entity will "ensure that they can 'pass' their own grading criteria."[280]

While determining which metrics to measure is an ongoing challenge for industry and regulators alike,[281] more thoughtful metrics are possible. For example, Facebook now reports a "prevalence" metric alongside its aggregated takedown numbers. Prevalence "estimates the percentage of times people see violating content" as opposed to its gross volume.[282] This metric incentivizes Facebook not only to remove more hate speech posts, but also to reduce how many times people *view* it — a more meaningful measure of the impact of its enforcement measures. This is indeed what happened: in an effort to decrease prevalence, Facebook made changes to "reduce problematic content in the News Feed."[283] That is, beyond just taking content down, Facebook made changes to the algorithm that determines how content is distributed. YouTube has also started reporting prevalence metrics and made them a key performance indicator for employees.[284] Beyond Facebook and YouTube, no other platform reports this metric and no legislative proposals would require them to do so.[285]

There is one other data point that is worth singling out as useful. Most platforms release separate reports about government requests to remove content.[286] These reports are an important indicator of the pressure governments place on platforms, but most are not comprehensive. They generally include only formal legal demands,[287] but governments also increasingly use informal mechanisms like "internet referral units"

[280] Sidney A. Shapiro & Rena Steinzor, *Capture, Accountability, and Regulatory Metrics*, 86 TEX. L. REV. 1741, 1744 (2008).

[281] Ben Whitelaw, *"It's a Challenge from a Morale Perspective": What It's Like Setting the Rules at Some of the World's Largest Platforms*, KINZEN (Sept. 29, 2021), https://www.kinzen.com/blog/trust-and-safety-report-challenges-success-metrics [https://perma.cc/LYB3-L6K9].

[282] Arcadiy Kantor, *Measuring Our Progress Combating Hate Speech*, FACEBOOK NEWSROOM (Nov. 19, 2020), https://about.fb.com/news/2020/11/measuring-progress-combating-hate-speech [https://perma.cc/NH82-VYAD].

[283] Rosen, *supra* note 81.

[284] Patel, *supra* note 83.

[285] There are competition tradeoffs with transparency mandates. More valuable metrics, like prevalence, would be more costly to produce and therefore could create barriers to entry for startups. Making the extensiveness and burden of transparency demands vary according to platforms' size would help. But anticompetitive impacts of requiring valuable metrics are not a good reason to require reporting of useless, but cheaper, ones.

[286] *The Transparency Reporting Toolkit: Content Takedown Reporting: Appendix*, NEW AM., https://www.newamerica.org/oti/reports/transparency-reporting-toolkit-content-takedown-reporting/appendix [https://perma.cc/K7SH-EE2L] (tracking content takedown reporting practices for a number of platforms).

[287] A notable exception is Twitter's reporting. Emma Llansó, *Twitter Transparency Report Shines a Light on Variety of Ways Governments Seek to Restrict Speech Online*, CTR. FOR DEMOCRACY & TECH. (May 4, 2017), https://cdt.org/insights/twitter-transparency-report-shines-a-light-on-variety-of-ways-governments-seek-to-restrict-speech-online [https://perma.cc/3GH4-5LF2].

that flag content for platforms to take down under the platforms' *own* rules rather than using formal legal orders.[288] When such demands are omitted from transparency reporting, the relationship between platforms and governments remains murky.[289] Unsurprisingly, perhaps, these separate reports about government requests are rarely required by regulation.

In sum, regulatory discussions of transparency mandates epitomize the blind spots this Article has described. They focus on a very narrow slice of content moderation (takedown/leave-up decisions) and do not take into account the tradeoffs behind this data, the perverse incentives they create, the influence of other actors outside the standard picture, and the need to develop a more responsive and continuous form of oversight. A one-size-fits-all transparency mandate will be too generic to reveal useful information and too static to keep pace with content moderation's constant evolution. Such a mandate would produce reporting on outdated features and practices, without shedding light on anything else. Platforms will encourage this limited window into their operations, and regulators will welcome the ease of drafting.[290] But simple mandates are no substitute for effective ones.[291] The mere fact that more effective mandates may be too complicated or fluid to codify does not justify costly measures that provide little value in their stead.[292]

This is not to doubt that transparency can have benefits. In the context of content moderation, transparency is useful to the extent that it achieves one of six related goals. First, it can correct vast information asymmetries between insiders and outsiders about platforms' enforcement of their public policies, which can inform the market and regulatory responses. Second, it can enable other institutional actors, such as the media, civil society, and governments, to respond to content and content moderation. For example, knowing the extent and type of misinformation, the identity of targeted populations, and the nature of platform interventions can help produce more effective counterspeech. Third, transparency can enable more effective stakeholder input into

[288] Chang, *supra* note 67, at 121–22.

[289] *See, e.g.*, The Lawfare Podcast, *Israel's "Cyber Unit" and Extra-legal Content Take-Downs*, LAWFARE (Apr. 29, 2021, 5:01 AM), https://www.lawfareblog.com/lawfare-podcast-israels-cyber-unit-and-extra-legal-content-take-downs [https://perma.cc/7Z3C-EXEM]; Evelyn Douek, *It's Not Over. The Oversight Board's Trump Decision Is Just the Start.*, LAWFARE (May 5, 2021, 3:11 PM), https://www.lawfareblog.com/its-not-over-oversight-boards-trump-decision-just-start [https://perma.cc/6UU6-WDL7]. It's worth noting that, even with these limitations, platforms are generally more transparent than governments about the number of these requests. Governments themselves should also release more detailed reporting about their requests to platforms to censor content.

[290] *See* BEN-SHAHAR & SCHNEIDER, *supra* note 249, at 5.

[291] *See id.* at 11 ("[B]ad law drives out good: mandates spare lawmakers the struggle of enacting better but less popular reforms.").

[292] *Id.* at 12.

content moderation design. Fourth, knowing the risks and harms of online content and content moderation could, in the future, provide a predicate for more coercive laws by providing an empirical basis for a substantial or compelling governmental interest in regulation. Fifth, greater transparency can, over time, coalesce into industry benchmarks of reasonable and responsible company behavior.[293] Finally, transparency can bring legitimacy and accountability to platforms' decisions.

But the point is that not all transparency mandates serve these goals. They are also not costless. The standard picture suggests transparency mandates that are simple, but writing effective mandates is anything but.

4. Process Theater. — The standard picture leads to content moderation regulation and discourse that generally take a rigid view of what process should be afforded to users — something akin to traditional judicial review, with an opportunity for affected users to be heard, receive individualized reasons for their treatment, and exercise a right to appeal.[294] This stems from an understanding of what is necessary to protect speech rights from infringement by *governments*, where such maximalist process is often cast as nonnegotiable: "Human rights law . . . requires more than just optimizing a speech-regulation system for a small quantity of error; it requires individualized determinations by independent arbiters."[295] Early content moderation scholarship focused on the threat state pressure posed to online freedom of expression,[296] and so this government-centric conception of the process that users should be afforded in every case may have seemed natural. But this focus persists in discussion about how *private* platforms should approach their voluntary moderation.[297] The assumption that users should be afforded an extensive set of procedural rights has largely gone unquestioned, including by regulators. This is partly because lawyers have often dominated platform regulation debates and "if all you've got is a lawyer, everything looks like a procedural problem."[298]

But copy-pasting a set of procedures from offline speech adjudication to online speech governance does not acknowledge the significant differences between the two systems or the fact that content moderation is conducted by private, not state, actors. It also assumes that more

[293] *See infra* section IV.C, pp. 603–06.

[294] *See* Yuval Eylon & Alon Harel, Essay, *The Right to Judicial Review*, 92 VA. L. REV. 991, 997 (2006); *see also, e.g.*, *DSA*, *supra* note 216, arts. 16, 17 & 20; PACT Act, *supra* note 223; BRADFORD ET AL., *supra* note 22, at 33–39; *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211.

[295] Llansó, *supra* note 211, at 4.

[296] *See generally* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296 (2014); Bambauer, *supra* note 67.

[297] *See, e.g.*, *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211.

[298] Nicholas Bagley, *The Procedure Fetish*, 118 MICH. L. REV. 345, 380 (2019).

individual process is always better — which is not true of public *or* private governance.

Extensive individual procedural requirements are often fetishized.[299] But even in the context of governmental (not private) power, "procedural rules must always be designed as a system, in light of the overall goal of the . . . program[s]."[300] What process is due "does not turn on the result obtained in any individual case," but "the risk of error . . . as applied to the generality of cases."[301] The question must be "[w]hich set of procedures will best balance the competing goals of efficiency, the protection of legal rights, and public accountability?"[302] This is a question of "mass, not individual justice."[303] If process is not "systems-rather than case-oriented, it will be irrelevant."[304]

The futility of maximalist individual process can be illustrated by taking the argument for individual due process to its extreme: imagine a mandate that every time a piece of content is taken down or flagged by a user, a human must review the case, give the user an opportunity to be heard, provide reasons for their ultimate decision, and offer the chance to appeal.[305] This would be an absurd system, divorced from the reality of platform scale. It would be entirely impractical and result in most users not receiving any process at all given the percentage of claims that could be resolved would be miniscule. It would incentivize platforms to shirk by adopting strategies like making appeals (even more) perfunctory. The length of time it would take to get to a final decision would drag out. The question is always: is "procedure for *procedure's own sake* the more important value — even if upholding that value means fewer cases get resolved . . . ?"[306] The fact that no one argues for this level of process (although the EU's draft DSA, as described above, gets close[307]) is implicit recognition of the need to evaluate due process systemically.[308] And the fact that due process mandates often include carveouts for spam is further acknowledgement of the need to balance the costs and benefits of additional procedure in specific

[299] *Id.* at 400.

[300] Adrian Vermeule, *Deference and Due Process*, 129 HARV. L. REV. 1890, 1903–04 (2016) (emphasis omitted).

[301] Walters v. Nat'l Ass'n of Radiation Survivors, 473 U.S. 305, 321 (1985) (quoting Mathews v. Eldridge, 424 U.S. 319, 344 (1976)); *see also* Ames et al., *supra* note 197, at 22.

[302] Bagley, *supra* note 298, at 352.

[303] JERRY L. MASHAW, DUE PROCESS IN THE ADMINISTRATIVE STATE 36 (1985).

[304] *Id.*

[305] This is, in fact, not far removed from some of the mandates being enacted or proposed in a number of regulatory models and advocated for in scholarly and civil society work. *See, e.g.*, sources cited *supra* notes 219–25; *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211.

[306] Gluck, *supra* note 201, at 1680.

[307] *See supra* p. 566.

[308] LOUIS KAPLOW & STEVEN SHAVELL, FAIRNESS VERSUS WELFARE 252 (2006) ("[W]e suspect that few fairness proponents actually hold absolutist views.").

contexts that underpins current regulatory and academic debates. This balancing should be made more explicit and considered, however.

It also cannot just be assumed that more individual process increases fairness. There is no single conception of "fairness." Bureaucratic determination of disputes is, as Professor Jerry Mashaw puts it, "caught on the horns of a dilemma" between binding rules that constrain arbitrariness in keeping with the rule of law and individualized adjudication.[309] Mashaw made this observation in the context of his study of the Social Security Administration, which determines over three million claims a year with around ten thousand employees.[310] This "crushing" pace[311] is an enviable workload ratio compared with the scale of content moderation and the millions of decisions platforms make every day.

Because of all these variables, claims about what legitimacy any extra procedural measure would bring to content moderation cannot be made in the abstract.[312] The considerations in each context will be complex. The rest of this section illustrates this through the example of appeals.

It is largely assumed that more appeals (and, when AI has been used, appeals to humans) will necessarily increase overall accuracy and user satisfaction.[313] Offline, appeals are also often "lauded as a sort of cure-all for erroneous decisionmaking."[314] But the value-add of appeals is not so simple.

First, which decisions get appealed can be selective, "depend on factors wholly unrelated to accuracy," and can be reflective of neither actual error patterns nor systemic flaws.[315] Data from the Oversight Board bears this out. Appeals to the Board from people who have had their content taken down by Facebook vastly outnumber appeals from people who flagged content to the platform that was subsequently left up.[316] A likely explanation is that people are more invested in the fate of their own posts than someone else's. There's no evidence it's because Facebook's takedowns are orders of magnitude less accurate. Thus, appeal rates to the Board are likely the result of factors wholly unrelated to fairness or accuracy.

[309] JERRY L. MASHAW, REASONED ADMINISTRATION AND DEMOCRATIC LEGITIMACY 188 (2018).

[310] *Id.* at 186–87.

[311] Ames et al., *supra* note 197, at 4 (referencing appeals for veterans' benefits).

[312] Bagley, *supra* note 298, at 369.

[313] *See DSA*, *supra* note 216, art. 17(5); *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211.

[314] Ames et al., *supra* note 197, at 23.

[315] *Id.*; *see also* DANIEL E. HO ET AL., QUALITY ASSURANCE SYSTEMS IN AGENCY ADJUDICATION: EMERGING PRACTICES AND INSIGHTS 8 (2021).

[316] OVERSIGHT BD., OVERSIGHT BOARD TRANSPARENCY REPORTS — Q4 2020, Q1 & Q2 2021, at 49 (2021), https://oversightboard.com/attachment/987339525145573 [https://perma.cc/CCA7-ZKV7].

Second, individual appeals skew content moderation governance because the interests that get represented are individual ones, not societal ones — for example, human rights prosecutors may object to evidence of war crimes being removed from YouTube, but only the users on the ground in the middle of a crisis situation have the ability to appeal the relevant decisions because it is their content.[317] There is no mechanism for people not on platforms but affected by their decisions to have input. "Standing" limits the kinds of decisions that will be appealed.

Third, the assumption that appeals always increase accuracy does not always hold. Mandatory appeals to humans ignore that humans are not always more accurate than AI.[318] This, too, is a context-specific empirical claim that needs testing: humans often make mistakes, especially under severe time constraints.[319] As to legitimacy dividends, research suggests that people's preference for a human or AI decisionmaker is in fact highly contingent and sensitive to factors such as cost, speed, and error rates.[320] Human review has other costs too: automated review is cheaper, faster, and means fewer people are exposed to traumatic content.[321] As Huq argues, "[r]emedies for a due process deficit [in AI decisionmaking] are unlikely to take the form of additional human review but rather better algorithmic design."[322]

Finally, perhaps the most appealing argument for affording human review in every case is that it might increase a user's subjective feeling of procedural justice and therefore belief in the legitimacy of the process.[323] This is an important goal, but there are two caveats. First,

[317] O'Flaherty, *supra* note 107.

[318] Aziz Z. Huq, *A Right to a Human Decision*, 106 VA. L. REV. 611, 654 (2020). Facebook noted in its response to the Oversight Board's decision in Case Decision 2020-004-IG-UA that "automation can also be an important tool in re-reviewing content decisions since we typically launch automated removals only when they are at least as accurate as content reviewers." FACEBOOK RESPONSE TO OVERSIGHT BOARD DECISION 2020-004-IG-UA (Feb. 25, 2021), https://assets.documentcloud.org/documents/20491910/breast-cancer-response-full.pdf [https://perma.cc/K8LB-BQNN].

[319] SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA 173 (2019) ("The workers had just seconds to review and delete inappropriate material from user profiles."); Ames et al., *supra* note 197, at 18 ("The faster a decisionmaker has to work, the more she is likely to err.").

[320] Derek E. Bambauer & Michael Risch, *Worse Than Human?*, ARIZ. ST. L.J. (forthcoming) (manuscript at 3) (on file with the Harvard Law School Library).

[321] *See* Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, THE VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/A2KY-AN94].

[322] Huq, *supra* note 178, at 1906.

[323] BRADFORD ET AL., *supra* note 22, at 34–39.

again, this hypothesis needs testing. Does the average user in fact perceive a decision to be fairer if they know a human was involved?[324] Trust in human decisionmakers is also low. Furthermore, trust in humans versus algorithms may change over time, again highlighting the need for continuous and not static evaluation.[325] Blanket assumptions about what process leads to greater trust and satisfaction are just that — assumptions. Individuals' taste for procedural fairness in any given circumstance is "an entirely empirical issue."[326] Has there been a drop in trust following platforms' increased reliance on AI review during the pandemic, when they had fewer human reviewers available?[327] Not necessarily — if anything, platforms enjoyed a small reputational dividend for increasing the quantity and speed of takedowns.[328] Second, more fundamentally, it is important not to "conflate the short-term gain from human review . . . with the question of dynamic optimality."[329] Even if human review *does* increase an individual's satisfaction in *their* case, that is not the system's only goal. As Professor Nicholas Bagley suggests, additional procedure may increase trust in particular cases, but it also might impede a system's functioning in the aggregate, which ultimately decreases trust overall.[330] If procedural burdens cause platforms to simply fail to review content moderation decisions, or become too slow in doing so, this can hardly be described as a more effective system.[331]

Take Facebook's civil rights auditors' recommendation that all reports of voter interference be routed to a human moderator.[332] The company said this would not be "effective and efficient."[333] The company asserted that the vast majority of user reports don't violate its policies, so prioritizing individual review would "unintentionally slow

[324] *See, e.g.*, Reuben Binns et al., *"It's Reducing a Human Being to a Percentage"; Perceptions of Justice in Algorithmic Decisions* 9 (Ass'n for Computing Mach. Conf. on Hum. Factors in Computing Sys., Paper 377, 2018), https://dl.acm.org/doi/10.1145/3173574.3173951 [https://perma.cc/T7HV-MQZL] ("[W]hile algorithmic decision-making implicates dimensions of justice, it may also mitigate them . . . .").

[325] Wu, *supra* note 34, at 2023; Eugene Volokh, *Chief Justice Robots*, 68 DUKE L.J. 1135, 1170–71 (2019).

[326] KAPLOW & SHAVELL, *supra* note 308, at 12.

[327] Douek, *supra* note 77, at 802–03.

[328] Douek, *supra* note 78.

[329] Huq, *supra* note 318, at 663.

[330] Bagley, *supra* note 298, at 380.

[331] *See* Abbe R. Gluck et al., Essay, *Unorthodox Lawmaking, Unorthodox Rulemaking*, 115 COLUM. L. REV. 1789, 1842 (2015) (noting that a regulator deviating from regular procedure in the name of "getting its work done" can advance legitimacy (emphasis omitted)).

[332] MURPHY ET AL., *supra* note 241, at 33.

[333] FACEBOOK'S CIV. RTS. TEAM, FACEBOOK'S PROGRESS ON CIVIL RIGHTS AUDIT COMMITMENTS 28 (2021), https://about.fb.com/wp-content/uploads/2021/11/Metas-Progress-on-Civil-Rights-Audit-Commitments.pdf [https://perma.cc/2R39-GCTX].

[the] review process" without increasing overall accuracy.[334] These empirical claims by Facebook need verification, but there is no reason they couldn't hypothetically be true.

Blunt rules also create perverse incentives. The vast majority of content moderation decisions concern the application of platforms' own rules, and so can be changed at any time. Overly onerous procedural requirements can therefore incentivize companies to proscribe less speech to decrease enforcement burdens or make reporting mechanisms more difficult to use to reduce the number of posts flagged for review.[335] Facebook did exactly this when Germany introduced requirements in the NetzDG to have a separate reporting channel.[336] Twitter, by contrast, integrated the NetzDG reporting mechanism into its usual reporting process. As a result, it reports between 7236 and 26,215 times more complaints than Facebook.[337]

Procedural requirements must also be adaptable to exigencies, such as a surge in reports during periods of instability or emergency. Imagine, for example, requiring notice, opportunity to present a case, and reasons for every removal of the Christchurch Massacre video while it was going viral. Appeals should be prioritized based on the type of decision, the seriousness of the consequences, underlying error rates of initial decisions, and the potential for harm.

There are many more factors to consider. The type of decision matters: account suspensions warrant greater procedural protections than removals of content, for example, given the more significant ramifications for affected users, whereas a decision to label a post is a smaller burden on a user's expression than deletion and thus may justify lesser protection. Changes in technological capacity will also change the reasonableness of providing certain process. More AI decisionmaking may require more appeals, but if AI becomes more accurate, appeals might also be less useful. Platforms already make such tradeoffs and there are many kinds of decisions that users cannot appeal.[338]

A key difficulty for regulators is information asymmetries and uncertainty about what is technically possible. Platforms should not

[334] *Id.*

[335] MONIKA BICKERT, CHARTING A WAY FORWARD: ONLINE CONTENT REGULATION 9–11 (2020), https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf [https://perma.cc/MQ6L-U2SB].

[336] *Federal Office of Justice Issues Fine Against Facebook*, BUNDESAMT FÜR JUSTIZ (July 3, 2019), https://www.bundesjustizamt.de/DE/ServiceGSB/Presse/Pressemitteilungen/2019/20190702_1.html [https://perma.cc/NVX6-XWWT] (noting that the German Federal Office of Justice stated that Facebook's NetzDG reporting form was "too hidden").

[337] Ben Wagner et al., *Regulating Transparency? Facebook, Twitter and the German Network Enforcement Act*, FAT* '20, at 261, 267 (2020).

[338] These often include decisions regarding CSAM, spam, "behavioral" takedowns, other users' deletion of their comments, content labelling, or downranking decisions.

plead technological limitations to avoid responsibility to improve, but regulators should also not demand the technically impossible. This is why systemic transparency is important: it can mitigate this risk of complacency by creating cross-industry benchmarks for comparison. Ultimately, though, technological uncertainty can only be managed and not eliminated.

The underlying point is that mandating a particular kind of process in every case will freeze the current equilibrium in place, fail to engage with the competing equities that a content moderation system may reasonably be designed to achieve, disincentivize innovation, and create barriers to entry given the most well-resourced platforms will find it easiest to comply with such mandates. Decisions about the amount of process to afford must be made and defended systemically and dynamically, and the benefits of more process not merely assumed. In the end, "decision costs are impossible to evaluate normatively without understanding their sources and consequences," but there is definitely a point at which such costs become too high.[339] Simply transposing an "antiquated, unrealistic, and court-centric"[340] notion of due process from offline individual rights determinations offers the illusion of more accurate and accountable content moderation but not its reality.

* * *

Because designing a system of regulation to make content moderation "accountable" and "legitimate" is nebulous and progress is hard to measure, the subsidiary goals of mandating transparency and extensive procedural rights in paradigm cases have become the markers of a well-functioning regulatory system.[341] For platforms, maximizing individual removals or the availability of individual process becomes the marker of compliance, regardless of any impact on public welfare. Ex post review of individual decisions can be a valuable means of information gathering and quality assurance, but making this the raison d'être of content moderation regulation displaces the original goal of implementing oversight of content moderation *systems* with the goal of implementing oversight of *paradigm cases*.

This focus is a product of the ex post nature of content moderation accountability discourse. Once errors are identified, extensive procedures that may ensure *their* redress seem attractive even if they would make the overall functioning of the system more costly, slower, and more error prone.[342] If regulators take the standard picture as fixed, then

[339] Adam M. Samaha, *Undue Process*, 59 STAN. L. REV. 601, 620 (2006).

[340] Bagley, *supra* note 298, at 381.

[341] This is known as goal displacement. *See* John Bohte & Kenneth J. Meier, *Goal Displacement: Assessing the Motivation for Organizational Cheating*, 60 PUB. ADMIN. REV. 173, 174 (2000).

[342] *See* KAPLOW & SHAVELL, *supra* note 308, at 272–73.

individual procedural measures are a natural response to erroneous decisions. But the standard picture is neither comprehensive nor immutable: it omits much of the picture of content moderation and what *is* in frame is in constant flux. Relying on the standard picture of content moderation will produce simpler reforms, but it will give the illusion of progress without creating meaningful change. The desire to write simple rules is overtaking the importance of writing good ones.

IV. SECOND WAVE CONTENT MODERATION INSTITUTIONAL DESIGN

The previous Parts demonstrated the shortcomings of the current individualistic and ex post approach to the regulation of content moderation and why a more systemic and ex ante approach is necessary. This Part sets out the framework for such an approach to making content moderation systems more accountable to regulators and the public.

Some may query why the goal of regulatory reform of content moderation decisionmaking should be accountability. Accountability is, after all, a procedural, not substantive, regulatory goal.[343] To some, this may appear too thin a goal, given the significant political and social interests content moderation implicates. But accountability is not only an important minimum requirement for systems that have the broad societal impacts that content moderation systems do. It is also a pragmatic objective for regulatory reform. There are few points of agreement in political and academic debates about what the problems are with content moderation — or even if there are problems — let alone how best to fix them. But a common thread across the board in complaints about content moderation is distrust that companies enforce their rules in a way that is consistent with their public representations, whether from lack of will or ability. This distrust, and the desire to make platforms explain apparent disparities in their rule enforcement and how they will prevent any such disparities in the future, has prompted lawmakers around the world to, over the course of the last few years, hold hours of hearings and regularly make written demands for information from platform executives. The regulatory reforms outlined in this Part would channel those impulses into a structured and ongoing oversight regime. They would also force platforms to disclose the kind of information about their operations that would be necessary to craft more substantive regulatory reforms in the future. Accountability is therefore a necessary first step to any substantive intervention.

As in any regulatory system, making content moderation more accountable will create difficult questions about the optimal level of values

[343] *See supra* note 10.

like impartiality, transparency, and deliberation.[344] More is not always better. But the status quo with respect to content moderation is so lacking in any of these values that questions at the margins should not stall all reform. At the same time, as the previous Part argued, hasty reforms based on a limited and outdated picture of content moderation will not create the right forms of transparency or process and will leave many of the most consequential decisions and systems of content moderation out of frame and unaccounted for.

A systems thinking approach, by contrast, embraces rather than ignores the diversity and dynamism of content moderation systems, which are in constant flux. Instead of focusing on the downstream outcomes in individual cases, it focuses on the upstream choices about design and prioritization in content moderation that set the boundaries within which downstream paradigm cases can occur. Instead of creating barriers to entry or locking in a vision of content moderation that is fixed and reflects the practices of the current dominant firms, it would allow for innovation and iteration. And in focusing on procedural accountability rather than the pursuit of some substantive conception of an ideal speech environment, it is more politically feasible and less constitutionally vulnerable.

Some scholars and regulators have begun to reckon with the systemic nature of content moderation, but regulatory lag means the standard picture model still sits at the core of most regulatory and self-regulatory efforts to bring accountability to content moderation. This comes at the cost of finding more ambitious and imaginative solutions. This is particularly true in the United States,[345] perhaps because the dominance and magnetism of the First Amendment's individualistic understanding of speech makes this approach to speech governance especially hard to resist. To the extent that U.S. regulation has begun to reckon with the systems that sit behind paradigm cases, proposals are still piecemeal and do not deal with the full array of content moderation institutions and considerations. But such a blinkered model of content moderation regulation is particularly problematic in a legal system where the constraints of the First Amendment make the capacity for direct governmental regulation especially limited. To write regulations that bring

[344] ADRIAN VERMEULE, MECHANISMS OF DEMOCRACY: INSTITUTIONAL DESIGN WRIT SMALL 10–13 (2007).

[345] Daphne Keller, Opinion, *For Platform Regulation Congress Should Use a European Cheat Sheet*, THE HILL (Jan. 15, 2021, 1:00 PM), https://thehill.com/opinion/technology/534411-for-platform-regulation-congress-should-use-a-european-cheat-sheet [https://perma.cc/66UP-EXEF] ("The EU experience tells us a lot about pragmatic options to at least try to correct them. It would be ironic if the U.S., in its rush to change CDA 230, missed that lesson and embraced the same flawed rules that Europe is now abandoning.").

about more meaningful reforms, regulators need a broader understanding of the systems they seek to make accountable.[346]

The rest of this Part outlines a regulatory model based on such an understanding. The first bucket would require platforms to observe certain structural mandates that accommodate the different institutional arrangements in the content moderation industry but have two consistent aims. First, content moderation bureaucracies should be structured in a way that mitigates the possibility that bias or irrelevant considerations will prevent platforms from applying their rules as they publicly state they will. Second, the heterogenous decisionmakers involved in content moderation — currently absent from the standard picture — should be made identifiable so they can be held to account for their decisions.

The second bucket of reforms are procedural requirements directed at surfacing platforms' ex ante decisions about system design, creating processes for holding platforms to and testing their implementation of these commitments, and facilitating systemic transparency that can generate information for the industry and regulators, with a view to creating more specific standards and mandates in the future.

After describing these structural and procedural reforms, this Part then sets out a framework for their regulatory enforcement.

Together the reforms that follow can mitigate some of the most persistent concerns about content moderation's accountability deficits and produce a virtuous cycle of regulatory, public, and industry learning. The function of these processes is "not to control discretion but to make practice transparent. They facilitate diagnosis and improvement."[347]

A. Structural Mandates

1. Separation of Functions. — A pervasive concern about content moderation — perhaps the most universal and persistent — is that platforms pursue their own political and financial interests despite their public commitments to enforce their rules in a neutral manner. Rather than relying on ad hoc ex post individualistic review to surface and correct all such biases in enforcement, regulators should mandate structural separations that aim to "eliminate the incentives that would make [biased] conduct possible or likely in the first place."[348]

[346] Robert Gorwa, *What Is Platform Governance?*, 22 INFO. COMMC'N & SOC'Y 854, 865 (2019).

[347] Charles F. Sabel & William H. Simon, *Minimalism and Experimentalism in the Administrative State*, 100 GEO. L.J. 53, 80 (2011).

[348] Lina M. Khan, *The Separation of Platforms and Commerce*, 119 COLUM. L. REV. 973, 980 (2019).

Like administrative agencies,[349] platforms play multiple roles all at the same time — businesses, rule-writers and rule-enforcers[350] — but while administrative law requires separations of functions to mitigate concerns about bias, no such measures exist within platforms.[351] While the possibility of biased enforcement can never be completely eliminated, regulators can reduce opportunities for it to affect decisions. Platforms should be required to put a wall between those concerned with the enforcement of content moderation rules, on the one hand, and those whose job performance is measured against other metrics, such as product growth and political lobbying, on the other. This should be enforced through fines if the latter interfere with the decisionmaking of the former regarding individual content moderation decisions. This separation of personnel who have different functions and different incentives "is a fundamental principle of adjudication that is fair and perceived to be fair."[352] While keeping strict separations between different teams within the same company may have costs for efficiency, there are accuracy and procedural fairness gains. In the current context of deep and growing regulatory and public distrust that platforms are adhering to their public rules, the benefits outweigh the efficiency costs.[353]

The rule-enforcement personnel should report to a responsible designated officer who has high-level representation within the company and whose incentives are not affected by external considerations like product growth or advertising revenue. Putting rule enforcement in the hands of an identified actor with independent authority enhances accountability and fosters a sense of responsibility.[354]

Concerns about platforms' political biases are bipartisan and global. The purported anti-conservative bias of "Big Tech" has become a rallying cry in right-wing politics.[355] Facebook even commissioned a report

[349] *See, e.g.*, Peter L. Strauss, *The Place of Agencies in Government: Separation of Powers and the Fourth Branch*, 84 COLUM. L. REV. 573, 577–78 (1984).

[350] MACKINNON, *supra* note 22, at 154.

[351] 5 U.S.C. §§ 554(d), 557(d)(1). Of course, in administrative law, this applies only to formal adjudications, but the underlying principle and rationale can be applied more broadly, including to the informal private rulemaking and adjudication involved in content moderation. *See generally* Michaels, *supra* note 19 (advocating for a broad conception of separation of powers in the context of administrative law and privatized governance).

[352] ASIMOW, *supra* note 95, at 64.

[353] This may not be true when it comes to much smaller platforms that have resource constraints that require them to combine staffing functions. As such, like with other prescriptions in this section, the mandate could be drafted to kick in after a platform reaches a certain size.

[354] *See* Daphna Renan, *Pooling Powers*, 115 COLUM. L. REV. 211, 278 (2015) (citing Jacob E. Gersen, *Unbundled Powers*, 96 VA. L. REV. 301, 324–25 (2010)).

[355] *See, e.g.*, Letter from Ken Buck, Member of Cong., et al. to Members of the Facebook Oversight Bd. (Feb. 12, 2021), https://files.constantcontact.com/60ec52f3801/958b9753-5739-4a6a-

by a former Republican Senator to review "concerns regarding perceived bias by Facebook against conservative organizations, individuals, and lawmakers."[356] At the same time, others suspect platforms have made changes to intentionally *favor* conservatives.[357] Business interests also appear to influence where platforms dedicate resources globally.[358] Politics also often drives platforms' decisions in many places, from the United States to India.[359] As one employee told a reporter, "[t]oo often we've made politically expedient exceptions at the expense of our own rules, which we generally believe to be fair."[360] Or as former Facebook Chief Security Officer Alex Stamos put it, "[a] core problem at Facebook is that one policy org is responsible for both the rules of the platform and keeping governments happy."[361]

Bias concerns are not just political. The influence of advertisers on content moderation is understudied and opaque.[362] Platforms like to talk in the language of high principle when discussing their rules but they are profit-driven businesses.[363] As YouTube acknowledges, "[o]ur policies are designed [in part] . . . to make sure we are able to keep

8c7f-e3622960304a.pdf [https://perma.cc/DD2Y-N3ZG] (demonstrating concerns of Republican members of the House Judiciary Committee); Ashley Gold, *Republicans Raise Bias Claims to Board Reviewing Trump's Facebook Ban*, AXIOS (Feb. 11, 2021), https://www.axios.com/republicans-raise-bias-claims-to-board-reviewing-trumps-facebook-ban-164ec898-4569-45a1-bbc0-5708b5bcce1d.html [https://perma.cc/XC8L-D4PA].

[356] JON KYL, COVINGTON INTERIM REPORT (2019), https://about.fb.com/wp-content/uploads/2019/08/covington-interim-report-1.pdf [https://perma.cc/FHA4-LQM6]; *see also* Nick Clegg, *An Update on Senator Kyl's Review of Potential Anti-conservative Bias*, FACEBOOK NEWSROOM (Aug. 20, 2019), https://about.fb.com/news/2019/08/update-on-potential-anti-conservative-bias [https://perma.cc/D7CM-PF8K].

[357] *See, e.g.*, Monika Bauerlein & Clara Jeffery, *Facebook Manipulated the News You See to Appease Republicans, Insiders Say*, MOTHER JONES (Oct. 21, 2020), https://www.motherjones.com/media/2020/10/facebook-mother-jones [https://perma.cc/LK8Z-ZHYJ]; Craig Silverman & Ryan Mac, *Facebook Fired an Employee Who Collected Evidence of Right-Wing Pages Getting Preferential Treatment*, BUZZFEED NEWS (Aug. 6, 2020, 8:13 PM), https://www.buzzfeednews.com/article/craigsilverman/facebook-zuckerberg-what-if-trump-disputes-election-results [https://perma.cc/QY6Y-HTHB]; Stanley-Becker & Dwoskin, *supra* note 74.

[358] Craig Silverman et al., *"I Have Blood on My Hands": A Whistleblower Says Facebook Ignored Global Political Manipulation*, BUZZFEED NEWS (Sept. 14, 2020, 7:36 PM), https://www.buzzfeednews.com/article/craigsilverman/facebook-ignore-political-manipulation-whistleblower-memo [https://perma.cc/PD6A-WHEL]; CHLOE COLLIVER ET AL., HOODWINKED: COORDINATED INAUTHENTIC BEHAVIOUR ON FACEBOOK 5 (2020).

[359] Newley Purnell & Jeff Horwitz, *Facebook's Hate-Speech Rules Collide with Indian Politics*, WALL ST. J. (Aug. 14, 2020, 12:47 PM), https://www.wsj.com/articles/facebook-hate-speech-india-politics-muslim-hindu-modi-zuckerberg-11597423346 [https://perma.cc/ZKV8-JDND].

[360] Stanley-Becker & Dwoskin, *supra* note 74.

[361] Alex Stamos, TWTEXT.COM, https://twtext.com/article/1265394955515650050# [https://perma.cc/97D9-LQQL].

[362] *See* Waldman, *supra* note 79 (manuscript at 11).

[363] *See* Biz Stone, *The Tweets Must Flow*, TWITTER BLOG (Jan. 28, 2011), https://blog.twitter.com/en_us/a/2011/the-tweets-must-flow [https://perma.cc/WUZ2-VZM5]; *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/75XK-LNU9].

advertisers coming back to YouTube."[364] Platforms seem to be more responsive to advertisers' concerns than other stakeholders' — advertisers have managed to get platforms to commit to independent audits of their transparency reports where all others have failed.[365] Whatever one's priors about desirable speech rules, advertisers' preferences are unlikely to be a reliable proxy for the public good.[366] At the very least, accommodating advertisers' concerns belies platforms' repeated public statements that they enforce their rules impartially.[367]

No organizational fix can eliminate all bias. Platforms are businesses and will prioritize their bottom line. Individuals within them will always have their own professional and political motivations. But currently there is little to constrain even the most blatant biases despite platforms' outward insistence that they enforce their rules fairly. Ex post review fails when it can be interfered with by platform employees responsive to incentives other than the accurate and consistent enforcement of the rules. Having an organizational structure that represents platforms' public commitments to enforce their rules evenhandedly is the first step to realizing them.

A strength of the separation-of-functions principle is that it can accommodate the heterogeneity of content moderation institutions beyond the standard picture and operate upstream from individual decisions.[368] This adaptability makes it an especially attractive tool for regulating the diversity and fluidity of content moderation systems.

2. Complaints Mechanisms. — The standard picture relies on user-initiated complaints in individual cases to surface errors and has no

[364] Susan Wojcicki, *Letter from Susan: Our 2021 Priorities*, YOUTUBE OFF. BLOG (Jan. 26, 2021), https://blog.youtube/inside-youtube/letter-from-susan-our-2021-priorities [https://perma.cc/Y4D5-DUA6].

[365] Kate Kaye, *Getting Facebook, YouTube, TikTok, Twitter and Others to Independent GARM Brand Safety Verification Is a Diplomatic Dance*, DIGIDAY (May 24, 2021), https://digiday.com/marketing/as-facebook-commits-to-independent-garm-brand-safety-verification-getting-youtube-tiktok-twitter-and-others-on-board-is-a-diplomatic-dance [https://perma.cc/E9EF-P5BL].

[366] *See, e.g.*, Leon Yin & Aaron Sankin, *Google Blocks Advertisers from Targeting Black Lives Matter YouTube Videos*, THE MARKUP (Apr. 9, 2021, 8:00 AM), https://themarkup.org/google-the-giant/2021/04/09/google-blocks-advertisers-from-targeting-black-lives-matter-youtube-videos [https://perma.cc/HES6-PNHY].

[367] *See, e.g.*, 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/53WG-3LVJ] ("The goal of our Community Standards is to create a place for expression and give people voice. . . . We work hard to . . . apply[] our policies consistently and fairly to people and their expression."); *How YouTube Develops Policies*, YOUTUBE (May 10, 2021), https://www.youtube.com/watch?v=3A-MD13TQNE [https://perma.cc/E8XH-94EF] (describing how YouTube tries to ensure that every "policy can be consistently enforced by our thousands of different policy reviewers across the world"); Twitter, Inc., *World Leaders on Twitter: Principles & Approach*, TWITTER BLOG (Oct. 15, 2019), https://blog.twitter.com/en_us/topics/company/2019/worldleaders2019.html [https://perma.cc/H48R-SMFL] ("Our goal is to enforce our rules judiciously and impartially.").

[368] Khan, *supra* note 348, at 1063 ("A final functional justification for structural separations is that they are highly administrable.").

other mechanisms for surfacing systemic irregularities. Section IV.B below outlines a number of procedural requirements that are intended to incentivize due diligence and provide ongoing oversight of content moderation systems ("police patrols"), but there should also be a mechanism to sound fire alarms where such due diligence fails.[369] Media stories based on insider leaks or complaints by affected users constitute a rough, informal type of "fire alarm."[370] This Article has argued in part that the debate about content moderation is distorted by outsized reliance on such fire alarms that have limited ability to contextualize the complaints or engage in follow-up formal investigations. But as part of a more comprehensive governance framework, an external channel to raise red flags about particular breakdowns remains an important way of drawing attention to otherwise overlooked problems. The impact the Facebook Papers have had in focusing regulatory attention is evidence of the potential power of information from employees about poor system design.[371]

The regulatory body charged with overseeing the reforms in this Part should also have a channel for fielding complaints and the power to conduct investigations based on credible allegations of material misrepresentations in the disclosure requirements outlined below. This is both a primary oversight mechanism in the administrative state[372] and a standard form of consumer protection in private industries where there is asymmetric access to information and the possibility of "exit" is unlikely to sufficiently protect consumer interests.[373]

Because regulating speech is not like regulating physical substances like pollutants, regulators cannot field substantive complaints about lawful content on platforms (such as "there is too much hate speech

[369] Mathew D. McCubbins & Thomas Schwartz, *Congressional Oversight Overlooked: Police Patrols Versus Fire Alarms*, 28 AM. J. POL. SCI. 165, 166 (1984).

[370] Ryan Mac has an impressive track record of obtaining accounts of internal Facebook meetings. *See, e.g.*, Ryan Mac, *Amid Israeli-Palestinian Violence, Facebook Employees Are Accusing Their Company of Bias Against Arabs and Muslims*, BUZZFEED NEWS (May 27, 2021, 5:56 PM), https://www.buzzfeednews.com/article/ryanmac/facebook-employees-bias-arabs-muslims-palestine [https://perma.cc/4LAL-CXWV]. A prominent example of a whistleblower is Sophie Zhang from Facebook. *See* Silverman et al., *supra* note 358. A recent example of the many stories based on the accounts of affected users is Elizabeth Dwoskin & Gerrit De Vynck, *Facebook's AI Treats Palestinian Activists Like It Treats American Black Activists. It Blocks Them.*, WASH. POST (May 28, 2021, 8:09 PM), https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship [https://perma.cc/QX7Q-6RKG].

[371] *See* Cat Zakrzewski & Reed Albergotti, *The Education of Frances Haugen: How the Facebook Whistleblower Learned to Use Data as a Weapon from Years in Tech*, WASH. POST (Oct. 11, 2021, 11:12 AM), https://www.washingtonpost.com/technology/2021/10/11/facebook-whistleblower-frances-haugen [https://perma.cc/PP46-6ZCV].

[372] Gillian E. Metzger, *The Interdependent Relationship Between Internal and External Separation of Powers*, 59 EMORY L.J. 423, 429 (2009); Neal Kumar Katyal, *Internal Separation of Powers: Checking Today's Most Dangerous Branch from Within*, 115 YALE L.J. 2314, 2346 (2006).

[373] Ian Harden, *Ombudsmen and Complaint-Handling*, *in* THE OXFORD HANDBOOK OF COMPARATIVE ADMINISTRATIVE LAW 773 (Peter Cane et al. eds., 2020).

or medical misinformation on this platform"). But complaints about a platforms' processes can surface important and useful information. Investigations showing that platforms either do not have adequate systems in place for enforcing their own rules or are subverting those rules when convenient can become focal points for pressure both inside and outside a company.[374] Such problems will not be apparent in the ordinary ex post review of a content moderation decision, which focuses only on the application of a particular rule in a particular case.

3. Disclosure of Nature and Extent of Contacts with Third-Party Decisionmakers. — As described above, content moderation increasingly involves decisionmakers outside a platform,[375] from fact-checkers, to government agencies, to other platforms.[376] Current regulatory models, based on the standard model that does not incorporate these decisionmaking arrangements, would not, for the most part, do anything to render the nature of these relationships more transparent or accountable.

But there cannot be oversight of a system without knowledge of all its constituent parts. Therefore, a more holistic approach to platform governance would require platforms to disclose the nature and extent of involvement of outside decisionmakers in their content moderation and impose fines on companies that do not disclose when outside parties directly influence individual content moderation decisions. For example, how fact-checkers' ratings are incorporated into the platforms' decisionmaking process should be transparent. TikTok should not take down videos based on third-party fact-checking without disclosing when and how often this occurs. Governments should not have special reporting channels to flag posts to be taken down to platforms without the frequency and basis of such informal orders appearing in platforms' transparency reporting.

[374] *See, e.g.*, Margo Schlanger, *Offices of Goodness: Influence Without Authority in Federal Agencies*, 36 CARDOZO L. REV. 53, 55 (2014) (describing "Offices of Goodness" within administrative agencies as a mechanism to promote certain values); Shirin Sinnar, *Protecting Rights from Within? Inspectors General and National Security Oversight*, 65 STAN. L. REV. 1027 (2013) (reviewing the strengths and limitations of Inspectors General offices); *see also* Colin Scott, *Implementation: Facilitating and Overseeing Public Services at Street Level*, *in* THE OXFORD HANDBOOK OF COMPARATIVE ADMINISTRATIVE LAW, *supra* note 124, at 595, 608 ("[T]he norms overseen and developed by ombudsman schemes are increasingly seen as proactive and regulatory in character, intended not only to provide a basis for complaint, but also to shape ground-level administrative actions.").

[375] *See supra* sections II.A.2–3, pp. 542–45.

[376] In administrative adjudication, best practice would be to prohibit ex parte communication between outsiders and decisionmakers, ASIMOW, *supra* note 95, at 65, but such an approach would be impractical for content moderation where third parties are deeply embedded in content moderation decisionmaking and necessary to give decisions relevant expertise and legitimacy.

An ex post and individualistic approach to surfacing the extent of third-party involvement in a platform's content moderation will inevitably fail to reveal the full extent of these relationships. At best, such review will reveal third-party involvement in the limited, ad hoc, and arbitrary set of cases that happen to be appealed by users. To get a full picture of external parties' involvement in content moderation, proactive reporting about the nature of their input on a systemic basis is necessary, and not merely in individual cases.

4. Retention of and Provision for Access to Data. — Platforms increasingly voluntarily disclose the kind of information that the standard picture suggests is most important to making content moderation transparent and accountable by releasing regular reports including the number of takedowns, appeal and reversal rates, and how much violating content was detected by AI versus how much was flagged by users.[377]

But platforms are much slower to voluntarily release the kind of data that systemic understanding of content moderation highlights. Platforms rarely release information about the broader functioning of their systems. Platforms could, for example, report on the distribution of errors in their content moderation that might highlight weak areas in their enforcement capacities, or release data about the effectiveness of interventions other than taking posts down (such as reducing the distribution of or labeling certain posts), or how user behavior is affected by changing platform design and affordances (for example, making reporting mechanisms easier to use or increasing the number of emoji reactions users can give posts). It is unsurprising that platforms are reluctant to produce this data given it implicates more fundamental questions of platform design than the resolution of individual content moderation cases.

Researchers have long been calling for access to platform data that would help in understanding these broader impacts of social media and platforms' policies, a necessary prerequisite for informed legal and policy responses.[378] Despite four years of public and regulatory techlash,

[377] *See, e.g.*, *The Santa Clara Principles on Transparency and Accountability in Content Moderation*, *supra* note 211 (stating that these kinds of disclosures are the kind necessary to bring transparency and accountability to content moderation).

[378] *See, e.g.*, Nathaniel Persily & Joshua A. Tucker, *Conclusion: The Challenges and Opportunities for Social Media Research*, *in* SOCIAL MEDIA AND DEMOCRACY, *supra* note 272, at 313, 313; Ramya Krishnan & Alex Abdo, *How Do You Solve a Problem Like Facebook?*, KNIGHT FIRST AMEND. INST. (Oct. 14, 2021), https://knightcolumbia.org/blog/how-do-you-solve-a-problem-like-facebook [https://perma.cc/P3B7-42F8]; MATHIAS VERMEULEN, THE KEYS TO THE KINGDOM (July 27, 2021), https://s3.amazonaws.com/kfai-documents/documents/2e579e7afa/7.28.21-Vermeulen.pdf [https://perma.cc/LEA5-PBWE]; *The Disinformation Black Box: Researching Social Media Data: Hearing Before the Subcomm. on Investigations & Oversight of the H. Comm. on Sci., Space & Tech.*, 117th Cong. (2021) (statement of Laura Edelson, Co-Director of Cybersecurity for Democracy at New York University).

there is still little verifiable information about the problems and potential harms of content on social media platforms, how they manifest, and the effectiveness (or otherwise) of platforms' actions to address them: "For the most part, legislators are legislating in the dark — with faint light being cast by whistleblowers or well-spun public reports from the firms."[379] Independent research is necessary if policy responses are to be based in empirical reality. Others have suggested models for such access elsewhere, and it is beyond my scope to give a detailed description. The core component of any model, as in Professor Nathaniel Persily's draft legislation now introduced into Congress as the Platform Transparency and Accountability Act,[380] would be a conditional safe harbor for platforms from liability for violation of privacy or cybersecurity laws if they exercise reasonable care when providing access in accordance with such mandates.[381]

B. Procedural Requirements

Regulators should also use procedural requirements as information-forcing mechanisms to prompt platforms to publicly account for — and be more proactive in thinking about — how they will operationalize, and mitigate any risks to, the effective enforcement of their publicly stated rules.

1. Annual Content Moderation Plans and Compliance Reports. — Instead of replicating the standard picture's outsized reliance on ex post review of decisionmaking, lawmakers should require platforms to engage in ex ante planning and risk assessment. Platforms should publicly explain the purpose of their rules, how they will enforce them, and how they will guard against risks to such enforcement. These are procedural goals ("what systems do you have in place to achieve what you have publicly committed to?"), not substantive ones ("how will you achieve a positive speech environment?"). Tying risk assessments to substantive outcomes would be constitutionally suspect and empirically fraught (if the relationship between speech and material risks in the world were easily ascertained, much about content moderation regulation would be easier). But forcing platforms to account for the risks to the enforcement

[379] Nathaniel Persily & Joshua A. Tucker, *How to Fix Social Media? Start with Independent Research.*, BROOKINGS (Dec. 1, 2021), https://www.brookings.edu/research/how-to-fix-social-media-start-with-independent-research [https://perma.cc/KP7Y-VY5C].

[380] Tara Wright, *The Platform Transparency and Accountability Act: New Legislation Addresses Platform Data Secrecy*, STAN. CYBER POL'Y CTR. (Dec. 9, 2021), https://cyber.fsi.stanford.edu/news/platform-transparency-and-accountability-act-new-legislation-addresses-platform-data-secrecy [https://perma.cc/6TLH-L482].

[381] Persily has been campaigning for such access for a number of years. *See* Nathaniel Persily, *Facebook Hides Data Showing It Harms Users. Outside Scholars Need Access.*, WASH. POST (Oct. 5, 2021, 7:20 AM), https://www.washingtonpost.com/outlook/2021/10/05/facebook-research-data-haugen-congress-regulation [https://perma.cc/WK5P-2MJY].

of their rules and how they plan to address them can provide regulators with basic information they currently lack about the systems they seek to regulate.

Requiring planning and risk assessment is not a novel proposal. Management-based regulation of this kind is increasingly common outside the content moderation context.[382] Impact assessments were pioneered in the context of environmental impact assessments with the passage of the National Environmental Policy Act of 1969[383] (NEPA), and have since spread to many domains, from fiscal impact statements to human rights, data protection, privacy, food and industrial safety, and well beyond.[384]

But unlike many other areas where the government of the day can settle disagreement by simply proscribing a view in regulation (as for pollution or financial risk, for example), defining what platforms should plan *for* and assess the risk *of* is especially challenging (and in many instances likely unconstitutional).[385] There is no uniform "definition of 'impact' that can be simply operationalized"[386] in any context, but the problem is worse in the speech context. If "[a]ssessing environmental impacts is uncommonly difficult,"[387] to say there is no normative agreement on what would constitute "good" content moderation or what constitutes "risk" is an understatement. Empirical effects of speech regulation are deeply contested. Despite the increasing tendency to use environmental analogies to describe the impacts of toxic speech,[388] there

[382] *See generally* Cary Coglianese & David Lazer, *Management-Based Regulation: Prescribing Private Management to Achieve Public Goals*, 37 LAW & SOC'Y REV. 691 (2003) (analyzing management-based regulation in the context of food safety, chemical accident regulation, and pollution prevention).

[383] 42 U.S.C. §§ 4331–4347; Selbst, *supra* note 248, at 122; Ana Maria Esteves et al., *Social Impact Assessment: The State of the Art*, 30 IMPACT ASSESSMENT & PROJECT APPRAISAL 34, 34 (2012).

[384] Emanuel Moss et al., *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*, DATA & SOC'Y 10–11 (2021), https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf [https://perma.cc/M7JL-MXCV]; Coglianese & Lazer, *supra* note 382, at 696–700.

[385] This is a point underappreciated in many proposals for content moderation risk assessments or duty of care obligations, which do not specify exactly what risk or harm platforms should face liability for. *See, e.g.*, TOM WHEELER ET AL., SHORENSTEIN CTR., NEW DIGITAL REALITIES; NEW OVERSIGHT SOLUTIONS IN THE U.S.: THE CASE FOR A DIGITAL PLATFORM AGENCY AND A NEW APPROACH TO REGULATORY OVERSIGHT 16 (2020), https://shorensteincenter.org/wp-content/uploads/2020/08/New-Digital-Realities_August-2020.pdf [https://perma.cc/BFS9-WCGQ].

[386] Moss et al., *supra* note 384, at 7.

[387] SERGE TAYLOR, MAKING BUREAUCRACIES THINK: THE ENVIRONMENTAL IMPACT STATEMENT STRATEGY OF ADMINISTRATIVE REFORM 17 (1984).

[388] *See, e.g.*, JEREMY WALDRON, THE HARM IN HATE SPEECH 96 (2012); 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 [https://perma.cc/5SYF-BZS9]; Whitney Phillips, *The Toxins We Carry*, COLUM. JOURNALISM REV. (2019), https://www.cjr.org/special_report/truth-pollution-disinformation.php [https://perma.cc/JU2Z-G5XU].

are real limits to the comparison.[389] Even if it were possible to agree (which it most assuredly is not) that stamping out hate speech is a desirable end, agreeing what constitutes "hate speech" is not the same as identifying CO2.

But all platforms should be able to demonstrate that they have systems in place to enforce their public commitments with a baseline of consistency and accuracy. Through the process of developing a public plan, platforms "should be struggling to provide as articulately and coherently as [they] can for key risks [to their rule enforcement]."[390] Platforms should be able to answer what resources they have for content moderation, what monitoring systems are in place to ensure systems are operating as intended, and what feedback mechanisms exist to ensure continuous improvement and identification of vulnerabilities. Platforms should be able to describe and defend the goals of their systems, so that the values they are optimizing for when they make tradeoffs are apparent and they are made accountable for the ways in which they decide to pursue them. Every platform needs a plan because every platform must moderate.[391] Even platforms whose value proposition is to be "free speech" havens discover they must moderate to avoid driving users away or to mitigate legal risks.[392]

Requiring platforms to publish and explain plans for how they will enforce their own rules may sound like a feeble form of accountability. But it's hard to overstate both how ineffective platforms are at enforcing their rules, and how little is known about what systems they have in place to do so. Despite being a purely procedural (not outcome-based) form of accountability, requiring platforms to have publicly available plans for rule enforcement has four main benefits, which also distinguish this form of systems-based transparency from the transparency theater of aggregated information about individual cases.

First, requiring planning forces platforms to think proactively and methodically about potential operational risks.[393] The process of having

[389] The Supreme Court has highlighted the limits of a similar comparison to food purity regulation, noting that in the context of food regulation "the public interest in the purity of its food is so great as to warrant the imposition of the highest standard of care on distributors . . . but the constitutional guarantees of the freedom of speech and of the press stand in the way of imposing a similar requirement" on disseminators of speech. Smith v. California, 361 U.S. 147, 152–53 (1959).

[390] Simon, *supra* note 260, at 76.

[391] GILLESPIE, *supra* note 14, at 5.

[392] Mike Masnick, *Parler Speedruns the Content Moderation Learning Curve; Goes from "We Allow Everything" to "We're the Good Censors" in Days*, TECHDIRT (July 1, 2020, 10:43 AM), https://www.techdirt.com/2020/07/01/parler-speedruns-content-moderation-learning-curve-goes-we-allow-everything-to-were-good-censors-days [https://perma.cc/3UGJ-ZQX8]; David Gilbert, *QAnon Is Mad at Trumpworld Twitter Clone GETTR Because of All the Porn*, VICE NEWS (July 2, 2021, 8:13 AM), https://www.vice.com/en/article/5dbwgd/qanon-is-mad-at-trumpworld-twitter-clone-gettr-because-of-all-the-porn [https://perma.cc/SF34-7Z7E].

[393] *See* Selbst, *supra* note 248, at 122.

to articulate a plan itself engenders proactivity and highlights blind spots. Platforms are known for failure to anticipate key risks, so "Making [Platforms] Think"[394] is meaningful,[395] and a useful counterweight to the "Move Fast and Break Things" culture of Silicon Valley.

The difference between Facebook and YouTube's handling of the 2020 U.S. and Burmese elections illustrates these benefits. Due to reputational costs from previous controversies in those countries, Facebook invested significant resources in contingency planning for information incidents[396] and observers noted evidence of improvement in Facebook's enforcement of its rules.[397] By contrast, YouTube engaged in far more perfunctory planning. One YouTube executive told the *New York Times* that "the usual processes for making [content moderation] decisions will be sufficient."[398] In both the United States and Myanmar, the content moderation problems Facebook seemed to avoid arrived on YouTube's doorstep.[399] As one member of Burmese civil society put it, "When it comes to hate speech and disinformation in Myanmar, YouTube is the new frontier."[400]

Second, planning creates documentation of decisions and their rationales, facilitating future review and accountability.[401] The first step in holding companies to their promises is having a record of them. This record can then also inform future regulation and create leverage for public pressure.[402]

[394] *Cf.* TAYLOR, *supra* note 387.

[395] Yifat Nahmias & Maayan Perel, *The Oversight of Content Moderation by AI: Impact Assessments and Their Limitations*, 58 HARV. J. ON LEGIS. 145, 170 (2021) ("[A] key advantage of impact assessments is their ability to influence platforms' internal organizational conduct.").

[396] *Meeting the Unique Challenges of the 2020 Elections*, FACEBOOK NEWSROOM (June 26, 2020), https://about.fb.com/news/2020/06/meeting-unique-elections-challenges [https://perma.cc/W9FE-GVSP]; *Additional Steps to Protect Myanmar's 2020 Election*, FACEBOOK NEWSROOM (Sept. 23, 2020), https://about.fb.com/news/2020/09/additional-steps-to-protect-myanmars-2020-election [https://perma.cc/C9JL-ZLUL].

[397] Kevin Roose, *On Election Day, Facebook and Twitter Did Better by Making Their Products Worse*, N.Y. TIMES (Nov. 5, 2020), https://www.nytimes.com/2020/11/05/technology/facebook-twitter-election.html [https://perma.cc/P4M5-5RFJ]; Peter Guest, *Facebook's Experimental Hate-Speech Policy Seems to Be Working*, REST OF WORLD (Nov. 20, 2020), https://restofworld.org/2020/pressing-pause-on-fake-news-in-myanmar [https://perma.cc/PE8J-8FJA].

[398] Mike Isaac et al., *What to Expect from Facebook, Twitter and YouTube on Election Day*, N.Y. TIMES (Nov. 3, 2020), https://www.nytimes.com/2020/11/02/technology/facebook-twitter-youtube-election-day.html [https://perma.cc/78TG-FEDX].

[399] Casey Newton, *How YouTube Got Played on Election Day*, PLATFORMER (Nov. 4, 2020), https://www.platformer.news/p/how-youtube-got-played-on-election [https://perma.cc/825S-AQDL]; Fanny Potkin, *YouTube Faces Complaints of Lax Approach on Overseas Election Misinformation*, REUTERS (Dec. 18, 2020, 2:22 AM), https://www.reuters.com/article/us-youtube-myanmar-misinformation/youtube-faces-complaints-of-lax-approach-on-overseas-election-misinformation-idUSKBN28S0QE [https://perma.cc/66JG-H99A].

[400] Potkin, *supra* note 399.

[401] Selbst, *supra* note 248, at 122.

[402] Kaminski, *supra* note 176, at 1608–09.

Third, relatedly, transparent plans facilitate broader policy learning for regulators and across industry.[403] Comparative information would show industry best (or worst) practices.[404] There is, for example, little public information about how many human moderators each platform employs, their locations and languages, what automated tools they use, or even how much platforms spend on content moderation overall. With cross-industry reporting, over time certain practices may coalesce into more general compliance standards.[405] Platforms can also learn from each other as they begin to experiment with different approaches to content moderation. Does sticking a label on fact-checked information actually help users understand the accuracy of information?[406] What has been the impact of deamplification measures? When a platform introduces more friction into the process of sharing content, how does that impact the way content travels?[407] Platforms should not each have to reinvent the wheel as they adopt and adapt such measures for their own services.

Fourth, such transparency helps facilitate public involvement and comment.[408] Consultation with stakeholders is a form of checking and balancing,[409] and it is vital for building accountability and legitimacy.[410] Platforms have long been criticized for their poor stakeholder engagement, especially in developing countries. But despite scholarly consensus that public participation is crucial in impact assessments, there is no agreement on the form it should take.[411] Public comment processes can become engagement theater and a feeble attempt at legitimation.[412]

[403] Selbst, *supra* note 248, at 152 ("The second goal . . . is all about exporting knowledge to the public.").

[404] Kaminski, *supra* note 176, at 1605, 1608–09.

[405] *Id.*

[406] *See* Ben Kaiser et al., *Warnings that Work: Combating Misinformation Without Deplatforming*, LAWFARE (July 23, 2021, 2:30 PM), https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming [https://perma.cc/N64Q-JKHA].

[407] For an unusual example of transparency around a failed intervention, see Vijaya Gadde & Kayvon Beykpour, *An Update on Our Work Around the 2020 US Elections*, TWITTER: BLOG (Dec. 16, 2020), https://blog.twitter.com/en_us/topics/company/2020/2020-election-update [https://perma.cc/952V-U36A] ("We hoped this change would encourage thoughtful amplification . . . . However, we observed that prompting Quote Tweets didn't appear to increase context . . . .").

[408] *See* Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises, *Human Rights Impact Assessments — Resolving Key Methodological Questions*, ¶¶ 10–21, U.N. Doc. A/HRC/4/74 (Feb. 5, 2007), https://undocs.org/en/A/HRC/4/74 [https://perma.cc/Z97H-VKTM] (finding "[e]ngagement of human rights experts and local stakeholders is critical" to conducting a human rights impact assessment); *cf.* Anne N. Glucker et al., *Public Participation in Environmental Impact Assessment: Why, Who and How?*, 43 ENV'T IMPACT ASSESSMENT REV. 104, 104 (2013); Andrew D. Selbst, *Disparate Impact in Big Data Policing*, 52 GA. L. REV. 109, 178 (2017).

[409] PETER CANE, CONTROLLING ADMINISTRATIVE POWER 465 (2016).

[410] Adoption of Recommendations, 84 Fed. Reg. 2139, 2146 (Feb. 6, 2019).

[411] *Cf.* Glucker et al., *supra* note 408, at 109.

[412] Moss et al., *supra* note 384, at 22; *see* Esteves et al., *supra* note 383, at 38.

Nevertheless, the public desire to have input in content moderation is clear and all major platforms already engage in stakeholder engagement, to varying degrees.[413] Facebook publishes minutes of its "Product Policy Forum," where its policy team reviews internal and external input before making changes to the platform's rules.[414] TikTok has regional Advisory Councils of experts to consult on its policies.[415] The Facebook Oversight Board receives public comments on each of its cases and received over nine thousand in relation to the decision to suspend former President Trump's account.[416] Having an ongoing planning and review process would allow for this consultation to become more consistent, transparent, and prospective, with stakeholders engaging on an iterative basis at each review rather than only on an ad hoc basis.[417]

There have been tentative discussions of impact assessments in the content moderation context. The EU's proposed Digital Services Act would bring the model to content moderation by requiring the largest platforms to issue annual assessments that "diligently identify, analyse and assess any systemic risks in the Union stemming from the design or functioning of their service and its related systems, including algorithmic systems, or from the use made of their services."[418] But the DSA's definition of "systemic risks" illustrates the problem of tying planning to a substantive conception of impacts: such risks are defined nebulously as relating to dissemination of illegal content, exercise of fundamental rights, or potential manipulation of the platform with an actual or

[413] Brenda Dvoskin, *Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, VILL. L. REV. (forthcoming 2022) (manuscript at 13–15), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3986181 [https://perma.cc/RUN6-4EUC].

[414] *Product Policy Forum Minutes*, FACEBOOK (Nov. 15, 2018), https://about.fb.com/news/2018/11/content-standards-forum-minutes [https://perma.cc/69ZV-EWJV].

[415] Vanessa Pappas, *Introducing the TikTok Content Advisory Council*, TIKTOK (Mar. 18, 2020), https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council [https://perma.cc/K9T6-5LK4]; Arjun Narayan Bettadapur, *Introducing the TikTok Asia Pacific Safety Advisory Council*, TIKTOK (Sept. 21, 2020), https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council [https://perma.cc/S33A-DP3K]; Julie de Bailliencourt, *Meet TikTok's European Safety Advisory Council*, TIKTOK (Mar. 1, 2021), https://newsroom.tiktok.com/en-gb/tiktok-european-safety-advisory-council [https://perma.cc/XLU3-72PT].

[416] Cristiano Lima, *Facebook Oversight Board Swamped with Comments on Trump Case*, POLITICO (Feb. 11, 2021, 2:30 PM), https://www.politico.com/news/2021/02/11/facebook-oversight-trump-banned-468730 [https://perma.cc/WG4Z-FHWY]; *Case Decision 2021-001-FB-FBR*, OVERSIGHT BD. (May 5, 2021), https://www.oversightboard.com/decision/FB-691QAMHJ [https://perma.cc/U4JR-T3ZM].

[417] *See* Robin Kundis Craig & J.B. Ruhl, *Designing Administrative Law for Adaptive Management*, 67 VAND. L. REV. 1, 43 (2014) ("[F]ormulation of the plan itself is clearly an adaptive management moment that lends itself to public input.").

[418] *DSA*, *supra* note 216, art. 34.

foreseeable negative effect on things like "public health" and "civic discourse."[419] If it were possible to concretely quantify risks to "civic discourse" from online speech, content moderation would be much easier.

There have also been a few content moderation–related Human Rights Impact Assessments (HRIAs). Usually conducted by a third-party rather than the company itself, HRIAs are intended to identify and address a company's adverse effects on human rights.[420] But the limited track record of HRIAs in the content moderation context is more cautionary tale than inspiring precedent. Facebook commissioned HRIAs in four developing countries in 2018.[421] The completed HRIAs are cursory and general, revealing no new information about Facebook's internal operations or insight into the company's adverse effects in the relevant markets. The HRIA on Myanmar, for example, barely acknowledged the genocide in the country in which U.N. officials said social media, and Facebook in particular, played a "determining role."[422] The HRIAs also only made passing reference to Facebook's use of algorithms and automated moderation. One evaluation suggested these HRIAs might be "ethics washing."[423]

Critics have argued that impact assessments, like those mandated by NEPA, are ineffective because they lack substantive force.[424] Under NEPA, for example, once a risk assessment is completed, "the agency is free to simply ignore the problem and forge ahead"[425] and "the predictions contained in any given EIS [environmental impact statement] could turn out to be wildly inaccurate, and no one would be the

[419] *Id.* art. 26(1).

[420] *See* Human Rights Council Res. 17/4, U.N. Doc. A/HRC/RES/17/4, ¶ 1 (July 6, 2011); John Ruggie (Special Representative of the Sec'y-Gen. on the Issue of Hum. Rts. & Transnat'l Corps. & Other Bus. Enters.), *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, ¶ 15, U.N. Doc. A/HRC/17/31, annex (Mar. 21, 2011).

[421] Miranda Sissons & Alex Warofka, *An Update on Facebook's Human Rights Work in Asia and Around the World*, FACEBOOK (May 12, 2020), https://about.fb.com/news/2020/05/human-rights-work-in-asia [https://perma.cc/TE83-ZWHB]; Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, FACEBOOK (Nov. 5, 2018), https://about.fb.com/news/2018/11/myanmar-hria [https://perma.cc/8BDB-NQR6].

[422] Tom Miles, *U.N. Investigators Cite Facebook Role in Myanmar Crisis*, REUTERS (Mar. 12, 2018, 5:40 PM), https://www.reuters.com/article/us-myanmar-rohingya-facebook-idUSKCN1GO2PN [https://perma.cc/48KY-6KNF].

[423] MARK LATONERO & AAINA AGARWAL, HUMAN RIGHTS IMPACT ASSESSMENTS FOR AI: LEARNING FROM FACEBOOK'S FAILURE IN MYANMAR 1 (Mar. 19, 2021), https://carrcenter.hks.harvard.edu/files/cchr/files/210318-facebook-failure-in-myanmar.pdf [https://perma.cc/68VT-7HKC].

[424] Selbst, *supra* note 408, at 179–80; *see also* Strycker's Bay Neighborhood Council, Inc. v. Karlen, 444 U.S. 223, 227–28 (1980) (per curiam).

[425] Selbst, *supra* note 408, at 179.

wiser."[426] This point is well-taken, but in the context of content moderation, this weakness of a risk-assessment regime may be a strength. Avoiding substantive mandates is a precondition to most speech-related regulation, and purely process-based planning still brings benefits.[427] Indeed, it is precisely where potential impacts are hard to measure, desired substantive outcomes are difficult to define and impossible to prescribe, and mandating the use of particular technologies risks stifling experimentation in finding solutions, that a planning approach can be most useful.[428]

Nevertheless, without any follow-up, risk assessment can become theater no better than the transparency and due process mandates currently favored by regulators.[429] Content moderation plans so far have largely been of this nature — often the announcement of a plan has been the end of a platform's external engagement with an issue, rather than the beginning. For example, the public has been left almost entirely in the dark about the effectiveness of platforms' exceptional COVID-19 misinformation rules released to great fanfare. Two years after the adoption of the "Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online,"[430] there has been little public accounting of how companies have implemented their voluntary pledges. Therefore, any regulatory scheme must include an obligation for platforms to provide an annual public review of the implementation of their plans to create some measure of accountability for platforms' progress toward their goals.

Many would-be reformers may find requiring platforms to publish content moderation plans too feeble a requirement. But a slow start is typical for impact assessment regimes: "[T]hey develop over time . . . emerg[ing] and evolv[ing] from a mix of legislation, regulatory rulemaking, litigation, public input, and scholarship. . . . As precedents are established, standards around what constitutes an adequate account of impacts becomes stabilized."[431] There is, quite simply, no way of currently knowing what platforms have been doing, what works, and what doesn't. Appreciating this fact is a necessary first step to more sweeping reform.

[426] Bradley C. Karkkainen, *Toward a Smarter NEPA: Monitoring and Managing Government's Environmental Performance*, 102 COLUM. L. REV. 903, 927 (2002).

[427] Selbst, *supra* note 248, at 152–53.

[428] *Id.* at 123–24; *see* Coglianese & Lazer, *supra* note 382, at 701.

[429] *See, e.g.*, Ari Ezra Waldman, *Privacy Law's False Promise*, 97 WASH. U. L. REV. 773, 814 (2020) (making this argument in the context of privacy).

[430] *The Christchurch Call to Action: To Eliminate Terrorist and Violent Extremist Content Online*, CHRISTCHURCH CALL (May 15, 2019), https://www.christchurchcall.com/assets/Documents/Christchurch-Call-full-text-English.pdf [https://perma.cc/7W9T-JXHY].

[431] Moss et al., *supra* note 384, at 14.

2. Quality Assurance and Auditing. — Regulators should also require platforms to show they have quality assurance (QA) measures in place for their decisionmaking systems. QA is the "epitome of internal administrative law"[432] and a core requirement of agency oversight.[433] QA would seek to ensure that rule enforcement is relatively consistent and within a reasonable margin of error and "[c]ompared to other forms of error correction, such as episodic appeals, quality assurance programs may be better positioned to understand the source of adjudicators' error and the most effective interventions to remedy them."[434]

But what should QA measure in content moderation? "Quality" is itself a "deeply contested concept,"[435] and is not synonymous with accuracy.[436] Even if how rules should be applied were clear cut, is a 99% accuracy rate good enough?[437] On major platforms, this is still millions of mistakes. Should that 1% be false positives or false negatives? What if that 1% error rate disproportionately falls on marginalized groups? How much do the limits of technological capacity matter? If it were possible to achieve 99.5% accuracy but decisions would take an extra day, would that be an improvement in "quality"? Does the answer depend upon which kind of speech is involved — fraud, hate speech, defamation, child abuse material?

The only thing worse than trying to define "quality" is *not* trying. The task is infinitely harder with no insight into platforms' systems to draw on; with more data to compare across platforms and products, defining acceptable ranges can become possible.

The methodology and outputs of QA measures and platform transparency reporting should also be subject to independent auditing.[438] To be meaningful, disclosures need "to be backed with regulatory oversight," otherwise they can "obfuscate processes and practices beneath a veneer of respectability."[439] Without verification, given current information asymmetries, current transparency reports could be as accurate as Enron's financial statements, and no one would be the wiser.[440]

[432] Ames et al., *supra* note 197, at 29.

[433] *See* JERRY L. MASHAW, BUREAUCRATIC JUSTICE: MANAGING SOCIAL SECURITY DISABILITY CLAIMS 149 (1983); Ames et al., *supra* note 197, at 6–7.

[434] HO ET AL., *supra* note 315, at 6.

[435] Ames et al., *supra* note 197, at 47.

[436] *See* Jody Freeman, *Collaborative Governance in the Administrative State*, 45 UCLA L. REV. 1, 24 (1997).

[437] For comparison, in fiscal year 2018 the Board of Veterans' Appeals reported an accuracy rate of 93.6%. Ames et al., *supra* note 197, at 47.

[438] HO ET AL., *supra* note 315, at 24; Ames et al., *supra* note 197, at 70.

[439] Gorwa & Ash, *supra* note 272, at 291.

[440] *See* Khan, *supra* note 348, at 1034. *See generally* J. Nathan Matias et al., *Software-Supported Audits of Decision-Making Systems: Testing Google and Facebook's Political Advertising Policies*, PROC. ACM ON HUM.-COMPUT. INTERACTION, Apr. 2022.

3. Aggregated Claims. — In emphasizing the importance of ex ante accountability, the argument in this Article is not that *all* ex post review is ineffective. To the contrary: error review can be an important way to diagnose systemic failures. While the ex post review of individual cases that most current regulatory models favor will fail to bring meaningful systemic reform or accountability, review of aggregated claims could do exactly that.[441]

Instead of mandating additional appeals and procedural protections for individual users, regulators could instead require platforms to provide aggregated review mechanisms. Regulators should mandate that platforms review, as a class, all adverse decisions in a certain category of rule violation over a certain period (which can be varied to accommodate the different size and resources of the relevant platform).[442] Such a process is much more likely to identify institutional reform measures that could address system-wide failures or highlight trends and patterns. Aggregated claims are not merely a more efficient way of dealing with individual claims en masse but prompt a different kind of review, by directing attention to the roots of dysfunction.[443] This approach may also facilitate more transparent reasoning and justification because privacy concerns are less acute when discussing issues in the aggregate rather than in the context of particular cases.

Despite the benefits of aggregated claims and their widespread use in many areas of private law, their increasing application in the context of public rights has generated unease. Again, this may be why such an approach does not appear to have been considered at length in the content moderation context ("speech!" being such a sacred public right). But given the inefficiencies and ineffectiveness of addressing all claims individually, "[f]ar from undermining legitimate decision making, group procedures can form an integral part of public regulation and the adjudicatory process itself."[444] Noonan et al.'s extended comparison of public law litigation and bankruptcy litigation shows that it is unduly formalistic to dismiss such an approach in the public sphere because of amorphous concerns about "legitimacy."[445] The rationales for aggregated and structural interventions are the same in both public and private spheres. Arguing for such an approach for review of machine

[441] *See, e.g.*, Yoel Roth (@yoyoel), TWITTER (Jan. 17, 2022, 10:09 AM), https://twitter.com/yoyoel/status/1483094060554276867 [https://perma.cc/GM3A-VMEM] (statement of Twitter's Head of Site Integrity) ("We're already seeing clear benefits from reporting for (aggregate analysis) — especially when it comes to non-text-based misinfo, such as media and URLs linking to off-platform misinformation.").

[442] Van Loo, *supra* note 148, at 881 (discussing the different ways platforms may use aggregated claims).

[443] Ames et al., *supra* note 197, at 74–75.

[444] Sant'Ambrogio & Zimmerman, *supra* note 193, at 1645.

[445] Noonan et al., *supra* note 168, at 586.

learning decisions in the context of constitutional law, Huq notes that "[a]ggregate challenges . . . usefully direct attention to system-wide causes of constitutional harm. They invite remedies fashioned to account for the interests of all regulated subjects — and not, say, instruments that improve on accuracy for a subset of the regulated population while increasing errors for a majority."[446]

C. Enforcement

An underlying theme and motivation of this Article has been that the limits of direct governmental regulation of online speech are significant, making it necessary to find an approach that leverages and legitimates platform self-regulation. Governmental oversight of platforms should aim to maximize the private sector's resources, expertise, and dynamism in finding innovative and effective methods for tackling content moderation challenges while requiring platforms to explain, justify, and verify those methods. By allowing platforms to experiment, government oversight would avoid locking in the status quo at the major platforms.

Current regulatory approaches tend to take the structures platforms have already constructed to deal with content moderation as given. This is partly a sign of the success platforms have had in influencing regulatory and academic debates to reflect their own image. This Part has outlined a more fluid and less cramped vision for the structural and procedural reforms that are necessary. These are forms that are almost exclusively internal to platforms and could (and, in lieu of regulation, should) be implemented by platforms voluntarily. But governmental oversight will be needed to prompt reforms and build trust in their implementation.

For present purposes, the closest parallel to a regulatory system that leverages private self-regulation and transforms it into a public regulatory framework is in the realm of privacy regulation. Professors Daniel Solove and Woodrow Hartzog have documented how the Federal Trade Commission (FTC) has used its authority to police unfair and deceptive trade practices to develop a robust and substantive "Common Law of Privacy" jurisprudence.[447] The FTC began by enforcing companies' self-regulatory privacy policies, "serv[ing] as the backstop to the self-regulatory regime, providing it with oversight and enforcement — essentially, with enough teeth to give it legitimacy and ensure that people would view privacy policies as meaningful and trustworthy."[448] The

[446] Huq, *supra* note 178, at 1940.

[447] *See generally* Daniel J. Solove & Woodrow Hartzog, *The FTC and the New Common Law of Privacy*, 114 COLUM. L. REV. 583 (2014); Woodrow Hartzog & Daniel J. Solove, *The Scope and Potential of FTC Data Protection*, 83 GEO. WASH. L. REV. 2230 (2015).

[448] Solove & Hartzog, *supra* note 447, at 598–99.

FTC was confined to enforcing companies' own voluntary commitments because it lacked the power to enact substantive privacy rules and the United States has no other trans-sectoral privacy regulations.[449] These enforcement actions almost exclusively ended in settlement agreements. But over time, these agreements themselves developed into a body of best practices that set baseline industry standards and consumer expectations and now operate as a largely standalone set of substantive requirements.[450] Such requirements have included measures similar to those outlined in this Part, including "comprehensive privacy program[s]," in which companies commit to performing risk assessments, designating certain employees as responsible officers, and pursuing other measures tailored to their size and complexity.[451]

The EU's algorithmic accountability regime under the General Data Protection Regulation[452] (GDPR) also exemplifies an approach based on public accountability of private governance, as Kaminski documents.[453] Indeed, that regime also mirrors a number of the recommendations in this Article in requiring a data protection impact assessment that is a "continual process involv[ing] assessing risk, deploying risk-mitigation measures, documenting their efficacy through monitoring, and feeding that information back into the risk assessment."[454] This model of "monitored self-regulation" is more dynamic, better at leveraging the particular capacities of private and public sector actors, and can create a virtuous cycle of continuous improvement that changes internal company processes by making those companies consider and justify their approach to safeguarding public interests.[455]

The analogy is striking. In the content moderation context, the government is constitutionally prohibited from creating any comprehensive substantive speech rules. Even if it weren't, there would be persistent disagreement about what the substantive rules for content moderation *should* be. But there is more general agreement that the apparent mismatch between companies' content moderation policies on paper and their ability or willingness to enforce those policies fairly (or at all) is problematic. Holding companies to their speech rules is an important first step, just as holding companies to their privacy policies was. The requirements of structural reform and iterative public content moderation planning and review can provide indicators and evidence of systemic failures to enforce content moderation policies as written.

[449] *Id.* at 599.

[450] *Id.* at 649, 672.

[451] *Id.* at 617–18, 647 n.317.

[452] Council Regulation 2016/679, 2016 O.J. (L 119) 1 (EU).

[453] *See* Kaminski, *supra* note 176, at 1583.

[454] Margot E. Kaminski & Gianclaudio Malgieri, *Algorithmic Impact Assessments Under the GDPR: Producing Multi-layered Explanations*, 11 INT'L DATA PRIV. L. 125, 130 (2021).

[455] *Id.* at 131; *see id.* at 131–32.

Through a mechanism like the FTC's authority to police unfair and deceptive practices, a regulator could over time "add[] some teeth"[456] to platforms' content moderation commitments.[457]

An objection might be that because companies are under no obligation to make commitments to moderate, such a regulatory regime would have no substantive content. Faced with a regulatory framework they disliked, the argument goes, platforms would simply stop regulating speech on their platforms. In theory, this is a plausible argument. In practice, however, this objection falls away pretty quickly. The level of content moderation that platforms are legally required to perform is minimal, true. But this only proves the point: almost all existing content moderation is voluntary and yet there is plenty of it. Every platform must moderate to make its products attractive to its users.[458]

The absence of a core set of substantive requirements is thus both a strength and a weakness of this approach. Critics have argued that a regulatory approach like the one employed by the FTC does not provide adequate guidance to those it regulates of what is required to avoid liability.[459] But certainty is an instrumental value and should not be pursued at the cost of effective regulation. Content moderation, like data security, "changes too quickly and is far too dependent upon context to be reduced to a one-size-fits-all checklist."[460] Mandates for individual ex post review may be clear, for example, but they will also clearly fail to achieve their aims. The lack of prescriptive requirements allows regulation to adapt to each platform's individual context and to fluidly develop over time.

Perhaps the most obvious roadblock to the approach outlined here is a lack of regulatory capacity. The FTC's resources are already stretched and overseeing platforms' content moderation commitments will be no minor task. Whether platform oversight is given to the FTC or a new regulatory agency,[461] significant public investment will be necessary to make the idea of regulatory oversight a reality.[462] While Congress has been threatening platform regulation for years now, nothing has come of it. But this is another virtue of the substance-agnostic approach this Article has advocated: focusing on the structures and processes of content moderation makes political agreement more feasible.

[456] Solove & Hartzog, *supra* note 447, at 604.

[457] Van Loo, *supra* note 176, at 1620 & n.337.

[458] GILLESPIE, *supra* note 14, at 5.

[459] Hartzog & Solove, *supra* note 447, at 2244–45.

[460] *Id.* at 2259.

[461] *See, e.g.*, Persily & Tucker, *supra* note 379 ("The relevant enforcement body could be the Federal Trade Commission, given that it has been out front in dealing both with fraud and user privacy, or a wholly new government agency.").

[462] Van Loo, *supra* note 176, at 1619 (observing in the context of tech platform oversight that "[m]onitoring requires personnel, so the FTC would need to either obtain new allocations or reassign existing employees").

Lawmakers on both sides of the aisle are angry at platforms, but for different reasons. The ire about the gap between platforms' public statements and their actual practices, however, is bipartisan.

CONCLUSION: MOVING SLOWLY AND FIXING THINGS

Consensus about the correct balance of tradeoffs involved in content moderation decisionmaking is not going to magically appear anytime soon, and any temporary equilibrium that is achieved will be constantly disrupted by new norms, technologies, and social and political contexts. The question to ask when thinking about how to regulate content moderation is therefore not "what is the correct balance between competing equities?" but "how can the social thinking necessary for intelligent tradeoffs between different goals be institutionalized?"[463] While there will never be agreement on what constitutes "good" content moderation, there is growing convergence around one thing: the status quo of private companies determining matters of such public significance without any form of accountability, transparency, or meaningful public input is inadequate.

This Article offers a path forward that is incremental and experimental. No system of accountability emerges fully formed like Athena from the head of Zeus. Impact assessments, for example, have evolved in the decades since NEPA was enacted through a mix of legislation and regulation, litigation, public input, and scholarship.[464] Norms and standards for auditors have changed over time to meet evolving expectations of what audits should do.[465] Gradually, practices coalesce into standardized methods and industry norms. Shedding light on the actual operation of these systems might only be an interim step toward a more prescriptive regime in the future, but it is a necessary one.[466]

For lawyers, the frame of individualistic, ex post error correction is a familiar format with which to engage in discussion about speech disputes. But governing content moderation by trying to regulate individual decisions is using a teaspoon to remove water from a sinking ship. A regulatory regime based around ex post individual error correction may be more certain and familiar, but it is also more certain to fail to achieve meaningful accountability of content moderation systems as a whole or encourage necessary innovation. Current debates and regulatory proposals fixate on a narrow slice of content moderation at a particular snapshot in time — the application of a fixed rule to a specific

[463] TAYLOR, *supra* note 387, at 3.

[464] Moss et al., *supra* note 384, at 14.

[465] MICHAEL POWER, THE AUDIT SOCIETY: RITUALS OF VERIFICATION 64 (1999); Bamberger, *supra* note 10, at 453 (describing "the importance of reexamining and experimenting with the role" of auditors).

[466] Selbst, *supra* note 248, at 168–69 (noting the same about algorithmic impact statements).

piece of content — and do not acknowledge the difficult design decisions and tradeoffs that must be made upstream of paradigm cases. Regulation based on limited understanding of the underlying content moderation systems it is trying to regulate will achieve a limited form of accountability. The second wave of content moderation institutional design must be based on systems thinking and a comprehensive understanding of the task involved. This Article has provided a more expansive picture of content moderation than that which underpins many proposed reforms, and has suggested tools that could be used to bring that picture into frame, and make its underlying systems more accountable.