OF SYSTEMS THINKING AND STRAW MEN

Kate Klonick[*]

INTRODUCTION

In *Content Moderation as Systems Thinking*,[1] Professor Evelyn Douek, as the title suggests, endorses an approach to the people, rules, and processes governing online speech as one not of anecdote and doctrine but of systems thinking.[2] She constructs this concept as a novel and superior understanding of the problems of online-speech governance as compared to those existent in what she calls the "standard [scholarly] picture of content moderation."[3] This standard picture of content moderation — which is roughly five years old[4] — is "outdated and incomplete," she argues.[5] It is preoccupied with anecdotal, high-profile adjudications in which platforms make the right or wrong decision to take down certain speech and not focused enough on the platform's design choices and invisible automated removal of content. It draws too heavily from First Amendment contexts, which leads to platforms assessing content moderation controversies as if they were individual judicial cases.[6]

Douek calls her approach "both ambitious and modest."[7] The modest part calls for structural and procedural regulatory reforms that center content moderation as "systems thinking."[8] The notion of systems thinking conveys a generalized approach of framing complexity as a whole comprised of dynamic relationships rather than the sum of segmented parts.[9] The ambitious part is dismantling the standard picture of content moderation scholarship and challenging the resultant "accountability theater" created by platforms and lawmakers alike.[10] In Douek's view, it is this "stylized picture of content moderation"[11] that is

---

[*] Associate Professor of Law, St. John's Law School.

[1] Evelyn Douek, *Content Moderation as Systems Thinking*, 136 HARV. L. REV. 526 (2022).

[2] *Id.* at 530.

[3] *Id.* at 530, 535.

[4] *See generally* TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598 (2018).

[5] Douek, *supra* note 1, at 538.

[6] *Id.* at 556, 563. This is the focus on "paradigm cases" of individual speech decisions that "ignore[] the ex ante design choices platforms make," *id.* at 545, and "assume[] the necessity of a model of speech governance and the judicial role adapted from the First Amendment context," *id.* at 539.

[7] *Id.* at 533.

[8] *See id.* at 585.

[9] *See generally* LUDWIG VON BERTALANFFY, GENERAL SYSTEM THEORY (1969).

[10] Douek, *supra* note 1, at 533.

[11] *Id.* at 528.

to blame for regulators assuming "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."[12]

There is much to like about understanding content moderation as a complex, dynamic, and ever-evolving system. Particularly useful for an article titled *Content Moderation as Systems Thinking* that calls for regulation of technology, there is rich and detailed scholarship on content moderation in both sociotechnical theory and the law. Indeed, most of the academic work on content moderation is done by sociotechnical theory scholars who study content moderation and platform governance using systems-thinking and systems-theory frameworks.[13] Sociotechnical systems theory posits that an organization is best understood and improved if all parts of the system — people, procedures, norms, culture, technology, infrastructure, and outcomes — are understood as relational and interdependent parts of a complex system.[14] In analyzing private law under this theoretical framework, Professor Henry Smith describes systems as "a collection of elements and — crucially — the connections between and among them; complex systems are ones in which the properties of the system as a whole are difficult to infer from the properties of the parts."[15] Examples of systems abound at all levels of nature and society: from cognition to social networks or economies, or as Smith proposes, systems of law.[16]

---

[12] *Id.* at 531.

[13] A nonexhaustive list of scholars studying content moderation in this context includes Sarah Myers West, Robyn Caplan, Robert Gorwa, James Meese, Edward Hurcombe, and Ysabel Gerrard and Professors Tarleton Gillespie, Sarah T. Roberts, Christian Katzenbach, Mike Ananny, Philip Michael Napoli, José van Dijck, Alice E. Marwick, Rachel Kuo, Thomas J. Billard, Rachel Moran, Thomas Poell, David B. Nieborg, André Brock, and Chelsea Peterson-Salahuddin.

[14] *See generally* H.J. Leavitt, *Applied Organizational Change in Industry: Structural, Technological, and Humanistic Approaches*, *in* HANDBOOK OF ORGANIZATIONS 1144 (James G. March ed., 1965); Albert Cherns, *The Principles of Sociotechnical Design*, 29 HUM. RELS. 783 (1976).

[15] Henry E. Smith, *Systems Theory*, *in* THE OXFORD HANDBOOK OF THE NEW PRIVATE LAW 143, 144 (Andrew S. Gold et al. eds., 2020) (citing MELANIE MITCHELL, COMPLEXITY: A GUIDED TOUR (2011)); HERBERT A. SIMON, THE SCIENCES OF THE ARTIFICIAL (2d ed. 1981)).

[16] *Id.* at 144–45. This framework is recognizable in other theories of law, though it is not always recognized as a systems theory approach per se. See, for example, the landmark work of Professor Kimberlé Crenshaw, *Demarginalizing the Intersection of Race and Sex: A Black Feminist Critique of Antidiscrimination Doctrine, Feminist Theory and Antiracist Politics*, 1989 U. CHI. LEGAL F. 139, 140, which draws a similar systemic theoretical picture. In this revolutionary article, Crenshaw argues that the harms associated with being a Black woman do not run "along a single categorical axis" of race or gender, "because the operative conceptions of race and sex become grounded in experiences that actually represent only a subset of a much more complex phenomenon." *Id.*

Systems *thinking*, then, according to those that study it, is one step removed: "literally, a system of thinking about systems."[17] This definition is, of course, tautological; even the authors of the only article Douek cites on the topic seem confused.[18] But the takeaway of "systems thinking" is much the same as that described by sociotechnical theory and by Smith: an "understanding of dynamic behavior, systems structure as a cause of that behavior, and the idea of seeing systems as wholes rather than parts" — wholes that create "emergent properties" whose origins cannot be traced to any one part or interplay of the system.[19] It is both the ocean *and* the wave, the forest *and* the trees, as well as all of the interactions and the emergent properties resultant.[20]

I would fully support and could barely disagree with such a holistic conception, especially in the context of global online speech controlled and governed by private platforms. But evaluating systems thinking as a concept is difficult because Douek never defines this new approach or engages with any of the relevant scholarship or literature, save for a single autological definition in a footnote.[21] Instead, *Content Moderation as Systems Thinking* attempts to distinguish itself from the "standard scholarly picture" in which content moderation is "a privatized hierarchical bureaucracy that applies legislative-style rules drafted by platform policymakers to individual cases and hears appeals from those decisions."[22] Unfortunately, however, the standard-picture model of content moderation scholarship outlined by Douek simply does not exist. None of the works that Douek cites to for this model ever describe content moderation in such reductionist terms. Rather, for over two decades, online speech scholars, myself included, have consistently

---

[17] Ross D. Arnold & Jon P. Wade, *A Definition of Systems Thinking: A Systems Approach*, 44 PROCEDIA COMPUT. SCI. 669, 670 (2015) (summarizing the common elements among the myriad definitions of "systems thinking" to "defin[e] systems thinking [through] the application of systems thinking to itself").

[18] *See id.* ("The term has been defined and redefined in many different ways since its coining by Barry Richmond in 1987. What makes systems thinking so difficult to define? Why is it constantly redefined? What is everyone missing? Perhaps, rooted in our own field, lies the answer to defining the elusive concept of systems thinking in a way that will allow it to be measured. To this end, proposed is a surprisingly straightforward step in defining systems thinking — the application of systems thinking *to itself*.").

[19] *Id.* at 674; *see also* Henry E. Smith, *The Ecology of the Common Law*, 9 BRIGHAM-KANNER PROP. RTS. J. 153, 157 (2020) ("These emergent properties stem from the interactions and preclude strong forms of reductionism. Studying a car by taking it apart and studying the parts in isolation will not tell us much about the functions served by cars or their subcomponents. Cars, like many systems, are complex but not unmanageably so. Complexity comes along a spectrum, running from simplicity — in which components contribute additively to the whole — to disorganized complexity or even chaos — where small changes at the micro level can lead to large and unpredictable changes at the macro level." (citation omitted)).

[20] Arnold & Wade, *supra* note 17, at 671 (describing the position of Barry Richmond, the originator of the systems thinking term, that systems thinking entails "people embracing Systems Thinking position themselves such that they can see both the forest and the trees; one eye on each").

[21] Douek, *supra* note 1, at 530.

[22] *Id.* at 535.

described private content moderation in the very same language as Douek offers: as "systems of governance"[23] leveraging automated[24] and "human"[25] "processes"[26] created by a "constellation of actors"[27] who design dynamically and react "iteratively"[28] to "internal and external influence," in which freedom of speech and the First Amendment are only nominally the issues.[29]

This misrepresentation has impact. Because Douek does not fully engage with the depth of the scholarship that has already explored the issues she discusses, the article misdiagnoses why policymakers and popular commentators have failed to take account of the full picture of content moderation — and who is to blame. It is not "regulatory lag" driven by a misleading "standard picture" from scholars.[30] Nor are discussion of First Amendment analogies or a focus on procedural due process solutions at fault for the woes or lack of regulation. By framing the future of online speech as a binary choice between old and new, Douek makes the future of online speech seem like an either-or scenario in which the "first wave" does it wrong, while a new "second wave" would get it

---

[23] Klonick, *supra* note 4, at 1599; *see* Jack M. Balkin, Essay, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2028–29 (2018); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 28 (2019); Rory Van Loo, *Federal Rules of Platform Procedure*, 88 U. CHI. L. REV. 829, 832 (2021).

[24] REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM 153–54 (2012); 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*, 7 BIG DATA & SOC'Y 1, 2 (2020); James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 63–70 (2015) (describing how moderation systems operate differently along several lines — automatic or manual, transparent or secret, ex ante or ex post, and centralized or decentralized); Klonick, *supra* note 4, at 1636–37.

[25] MACKINNON, *supra* note 24, at 154; Klonick, *supra* note 4, at 1638–40.

[26] DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 53 (2019); *see also id.* at 54–56; NICOLAS P. SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 8 (2019); Rory Van Loo, *The Corporation as Courthouse*, 33 YALE J. ON REGUL. 547, 559–60, 567 (2016).

[27] 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* 23 (Hans-Bredow-Institut, Working Paper No. 1, 2020) https://leibniz-hbi.de/uploads/media/default/cms/media/oww9814_AP_WiP001InsideFacebook.pdf [https://perma.cc/2LD4-WKW6]; *see also* Chinmayi Arun, Essay, *Facebook's Faces*, 135 HARV. L. REV. F. 236, 236 (2022) ("Facebook engages with states and publics through multiple parallel regulatory conversations, further complicated by the fact that Facebook itself is not a monolith. This Essay argues that Facebook has many faces — different teams working towards different goals, and engaging with different ministries, institutions, scholars, and civil society organizations. It is also internally complicated, with staff whose sympathies and powers vary and can be at odds with each other. Content moderation takes place within this ecosystem.").

[28] Klonick, *supra* note 4, at 1637, 1648.

[29] *See* MACKINNON, *supra* note 24, at 154; *see also* Daphne Keller, Amplification and Its Discontents 30–36 (June 8, 2021) (Knight First Amendment Inst. Occasional Papers), https://knightcolumbia.org/content/amplification-and-its-discontents [https://perma.cc/S2GV-2SMW]. *See generally* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD (2006).

[30] *Contra* Douek, *supra* note 1, at 585.

right.[31] This framing is not just evidentiarily incorrect, it exposes a serious logical flaw in the argument for a systems-theory approach. Even if the scholarship had overemphasized hierarchy and individual decisions, a systems-thinking approach would suggest that those parts would *still be* essential components of the very system of content moderation that Douek attempts to describe. The elements of content moderation scholarship she eschews would need to be as accurate and true as trees if one is to understand the system of the forest and the emergent properties of their interaction.

This either-or approach also threatens to undo the hard-won improvements in transparency and procedural protections that scholars and advocates have fought to put in place to protect user rights and global free expression. Rather than acknowledging the ways in which these *existing* accountability approaches complement a "systemic" solution to content moderation or the plethora of scholarly debate over government control of speech, Douek proposes a set of reforms that are in many cases rehashed from existing literature. On their own, these reforms are indeed modest. But the proposed means of enforcing them is not modest at all: government control through a new agency to oversee the most invisible parts of content moderation "with a view to creating more specific standards and mandates" for online speech.[32]

In this Response, I first detail what *Content Moderation as Systems Thinking* gets right about content moderation, as well as what its characterization of existing scholarship gets wrong. I then show why the fact that the article oversells its reframing of this area of scholarship matters not just as a matter of accuracy, but also because it undermines efforts to achieve the real-world accountability that Douek — and so many others — are ultimately after.

The challenges of governing online speech are indeed "systemic." But proposing viable solutions requires more than merely describing the challenges as such, as evidenced by the fact that so many scholars already have. It requires recursive and iterative examination of one's priors, engagement with empirical realities and scholarly theories, and exploration of markets and governments besides one's own. In short, fixing the problems of online speech requires the very type of systems thinking which Douek names but does not employ.

I. THE "STANDARD PICTURE" STRAW MAN

Douek starts her article by presenting what she believes to be the problem and its cause: "This Article's central claim is that *the standard picture's focus on the treatment of individual posts is misguided* and that

---

[31] *Id.* at 534.
[32] *Id.* at 586; *see id.* at 605.

the toolset for content moderation reform needs to be expanded beyond individual error correction."[33]

There are roughly five implicit and explicit arguments that Douek makes to support this central claim:

First, a "standard picture of content moderation" exists and is primarily a result of academic scholarship.[34]

Second, in this standard scholarly 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.'"[35] Content moderation is a "privatized hierarchical bureaucracy that applies legislative-style rules drafted by platform policymakers to individual cases and hears appeals from those decisions."[36]

Third, the scholarly standard picture is inaccurate because it has "blind spots" that it fails to acknowledge: "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."[37]

Fourth, the scholarly standard picture is also inaccurate because it "is pervaded by First Amendment analogies."[38] This mistaken assumption is exemplified by how "content moderation is almost singularly concerned with the binary decision to take down or leave up individual pieces of content"[39] — the "high-profile content moderation controversies" like Nancy Pelosi looking drunk, Donald Trump being banned from Twitter, users denying the Holocaust, or the like.[40]

Finally, this misleading and incomplete scholarly standard picture is what "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."[41]

This Part takes these five issues in turn. Section A addresses the first of these claims, which is a question of construction. The "standard picture of content moderation" is a term and concept created by Douek, who defines it in a footnote reference to just eight academic works.[42] What makes these eight articles and books exemplary of the standard picture is not clear; the footnote omits mention of huge amounts of relevant influential scholarship and never provides reasoning or

---

[33] *Id.* at 530 (emphasis added).

[34] *See id.* at 535–39.

[35] *Id.* at 529 (quoting Klonick, *supra* note 4, at 1647).

[36] *Id.* at 535.

[37] *Id.* at 539.

[38] *Id.* at 556.

[39] *Id.* at 565.

[40] *Id.* at 536. It is unclear how Douek moves from "media headlines" to "scholarship" as the driving force behind inadequate reform, but that question will be discussed at the end.

[41] *Id.* at 531.

[42] *See id.* at 535 n.22.

methodology to explain its construction. Section B addresses the second, third, and fourth claims, which are substantive. Douek quotes narrowly from the literature she cites for the standard picture, and undercredits the works as a result. Section C addresses the fifth part of Douek's claim, which is causal. Douek does not adequately support the claim that the scholarly standard picture of content moderation is at fault for lawmakers' faulty attempts at regulation. Indeed, as Douek seems to recognize, blame for lawmakers' preoccupation with individual "high-profile" content moderation controversies is better placed on the media or lawmakers themselves.[43]

A. Constructing the Standard Picture

In the beginning of *Content Moderation as Systems Thinking*, Douek introduces the "standard picture" of content moderation scholarship.[44] Though she acknowledges it is "by no means [a] comprehensive" list, her citation references only eight scholarly works.[45]

Why reference these eight pieces — and not any of the hundreds of other books and articles published in the last decade on content moderation? It is not clear. The years of publication of the books and articles Douek cites range from 2012 to 2021, and many other articles and books on content moderation were published in the same window — as well as in the decade before. The cited pieces vary across disciplines, ranging from political science and communications studies books to law review and social science articles. They also vary in measures of apparent influence (partial as these measures may be). Several have been widely cited and downloaded, including Rebecca MacKinnon's *Consent of the Networked*,[46] Professor Tarleton Gillespie's *Custodians of the Internet*,[47] Professor David Kaye's *Speech Police*,[48] and my own *The New Governors*.[49] Others have only a few peer citations and fewer than

---

[43] *Id.* at 536.

[44] *See id.* at 535 & n.22 (citing Klonick, *supra* note 4, at 1639–41; GILLESPIE, *supra* note 4, at 116; Kyle Langvardt, *Can the First Amendment Scale?*, 1 J. FREE SPEECH L. 273, 298 (2021); Farzaneh Badiei et al., *Community Vitality as a Theory of Governance for Online Interaction*, 23 YALE J.L. & TECH. (SPECIAL ISSUE) 15, 33 (2021); Kettemann & Schulz, *supra* note 27, at 21–22; KAYE, *supra* note 26, at 53–57; MACKINNON, *supra* note 24, at 153–154; Marvin Ammori, *The "New" New York Times: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2276 (2014).

[45] Douek, *supra* note 1, at 535 n.22.

[46] Entry for *Consent of the Networked* by Rebecca MacKinnon, GOOGLE SCHOLAR, https://scholar.google.com/ [https://perma.cc/6FAN-WU9T] (select the "Articles" radio button and search "Consent of the Networked, MacKinnon" on Google Scholar) (869 citations).

[47] *Gillespie, Tarleton*, SCOPUS PREVIEW, https://www.scopus.com/authid/detail.uri?authorId=7102070921 [https://perma.cc/9YWG-DSQU] (707 citations).

[48] Entry for *Speech Police: The Global Struggle to Govern the Internet* by David Kaye, GOOGLE SCHOLAR, https://scholar.google.com/ [https://perma.cc/78Q3-CCKQ] (select the "Articles" radio button and search "Speech Police, David Kaye" on Google Scholar) (111 citations).

[49] *Klonick, Kate*, SCOPUS PREVIEW, https://www.scopus.com/authid/detail.uri?authorId=54784761500 [https://perma.cc/AT26-2DCA] (195 citations).

a hundred downloads.[50] Many of the cited pieces were written by scholars early in their careers or working at nonacademic institutions.[51] Yet many pieces that have been more frequently cited[52] or were written by

---

[50] For example, Kyle Langvardt's *Can the First Amendment Scale?* has just eighty-five downloads, *Can the First Amendment Scale?*, SSRN (Aug. 27, 2021), https://papers.ssrn.com/abstract=3911521 [https://perma.cc/XRF7-5X7N], and only six citations in other journals, *Kyle Langvardt, Can the First Amendment Scale?*, HEINONLINE, https://heinonline.org/HOL/P?h=hein.journals/jfspl1&i=273 [https://perma.cc/TR63-B36Y]. Matthias C. Kettemann & Wolfgang Schulz's article has twelve citations, Entry for *Setting Rules for 2.7 Billion: A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study* by Matthias Kettemann and Wolfgang Schulz, GOOGLE SCHOLAR, https://scholar.google.com/ [https://perma.cc/GY4C-25KP] (select the "Articles" radio button and search "Kettemann Schulz Rules for 2.7 Billion" on Google Scholar), while Farzaneh Badiei's co-written article has no citations, *Farzaneh Badiei et al., Community Vitality as a Theory of Governance for Online Interaction*, HEINONLINE, https://heinonline.org/HOL [https://perma.cc/GSU5-X4KE]. This is mentioned only to demonstrate that the selected scholarship in the standard picture is not uniform in its influence. It is not at all mentioned as insult to these authors, or to imply that their work is not excellent, valuable, or impactful.

[51] The authors Douek cites as representing the "standard picture of content moderation scholarship," Douek, *supra* note 1, at 535, include myself, an associate professor at St. John's Law School; Gillespie, a Senior Principal Researcher at Microsoft Research; Langvardt, assistant professor at Nebraska Law School; Kaye, clinical professor at UC Irvine Law School; MacKinnon, now in-house at Wikimedia Foundation; Ammori, chief legal officer at Uniswap; Kettemann & Schulz, scholars at Hans-Bredow, a small German think tank; and Badiei, now at Digital Medusa, a digital-governance advisory firm she founded.

[52] *See, e.g.*, DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014); ABRAHAM H. FOXMAN & CHRISTOPHER WOLF, VIRAL HATE: CONTAINING ITS SPREAD ON THE INTERNET (2013); SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019); SUZOR, *supra* note 26; SIVA VAIDHYANATHAN, ANTISOCIAL MEDIA: HOW FACEBOOK DISCONNECTS US AND UNDERMINES DEMOCRACY (2022); TIM WU, THE ATTENTION MERCHANTS: THE EPIC SCRAMBLE TO GET INSIDE OUR HEADS (2016); Jack M. Balkin, Commentary, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1 (2004) [hereinafter Balkin, *Digital Speech and Democratic Culture*]; Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296 (2014); Balkin, *supra* note 23; Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 PEPP. L. REV. 427 (2009); Bloch-Wehba, *supra* note 24, at 56; Robyn Caplan & Tarleton Gillespie, *Tiered Governance and Demonetization: The Shifting Terms of Labor and Compensation in the Platform Economy*, 6 SOC. MEDIA + SOC'Y 1 (2020); Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807 (2012); Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021); Gorwa et al., *supra* note 24; Grimmelmann, *supra* note 24; Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353 (2018); Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 NEW MEDIA & SOC'Y 4366 (2018); ROBYN CAPLAN, CONTENT OR CONTEXT MODERATION?, DATA & SOC'Y (2018), https://datasociety.net/wp-content/uploads/2018/11/DS_Content_or_Context_Moderation.pdf [https://perma.cc/WB3X-CHP2].

high-profile scholars in the field[53] are not referenced as comprising the standard picture.[54]

Indeed, the mystery of the scholarship that is left out from the standard picture is perhaps even more perplexing than what is left in. Douek cites to many of these additional scholarly sources in the second half of her article, but does so in *support* of her thesis, rather than as providing examples of the antagonist standard picture.[55] Much of this scholarship — the standard picture and its omissions — hardly differs in its descriptive or normative conclusions around content moderation.

There might be good reasons why Douek thinks these eight pieces of scholarship represent a "standard" picture while the scholarship she cites later does not. But absent any methodology or theory to explain it, the scholarship included and omitted from the standard picture is at best an arbitrary grouping.

B. Characterizing the Standard Picture

Douek's foundational argument is that the standard picture of content moderation scholarship has "blind spots" and "mistaken assumptions."[56] It is overly focused on "paradigm cases."[57] It fails to acknowledge that "[c]ontent moderation bureaucracies are a 'they' not an 'it' . . . made up of a sprawling array of actors and institutions, each of which has different functions and goals."[58] It neglects the "wide diversity of institutions . . . outside the hierarchical bureaucracy" of platform content moderation.[59] It ignores automatic "ex ante tradeoffs."[60] It "assumes the necessity of a model of speech governance and the judicial role adapted from the First Amendment context" and does not adequately grapple with the degree to which "[e]x [p]ost [r]eview [c]an [b]e [s]ystemic."[61]

But all of these things purportedly missing from the "standard picture" are in fact not missing at all. Indeed, the scholarly sources

---

[53] For example, compare the titles and affiliations of those included in the scholarly standard picture, *supra* note 51, with fields and titles of the scholars published in the same time frame and subject area but not included in the standard picture: Professor Jack M. Balkin, Yale Law School; Professor Julie Cohen, Georgetown University Law Center; Professor Genevieve Lakier, University of Chicago Law School; Professor Nathaniel Persily, Stanford Law School; Professor James Grimmelmann, Cornell Law School; and Professor Rory Van Loo, Boston University School of Law. *Cf.* Keerthana Nunna, W. Nicholson Price II & Jonathan Tietz, *Hierarchy, Race & Gender in Legal Scholarly Networks*, 75 STAN. L. REV. 71, 109 fig.3A, 110 fig.3B, 111 fig.4 (2023) (empirically showing how factors like school rank and hierarchy of scholarly networks impact acknowledgment citation).

[54] Section B will address the substance of these "standard picture omissions" in more detail.

[55] *See* Douek, *supra* note 1, at 556–64.

[56] *Id.* at 556.

[57] *Id.* at 535.

[58] *Id.* at 539.

[59] *Id.*

[60] *Id.*

[61] *Id.* at 539, 563.

cited in reference to the standard picture — and many others that go unmentioned — address these supposedly absent points, often multiple times and often in the very paragraphs and pages to which Douek cites. Moreover, many of these sources already describe content moderation in the *very terms* of systems theory.

As one example, take Douek's characterization of MacKinnon's 2012 book *Consent of the Networked*, a foundational 294-page study in internet policy and geopolitical power. Douek implies that MacKinnon misses the systems part of content moderation and thus is an example of the standard picture, summarizing MacKinnon's book in a footnote parenthetical as "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.'"[62] MacKinnon does in fact describe platform policy teams in this way, but here is the surrounding language around the quotation (emphasized for clarity) that Douek uses:

> Thus a big part of *the team's* job is to develop *processes* to identify abusive content and remove it, while not removing other postings or pages that may be edgy and upsetting to some but are not actually against the terms of service. They have developed a *system* that combines *automated* software to identify image patterns, keywords, and communication patterns that tend to accompany abusive speech, along with *review procedures* by flesh-and-blood *human* staff. Willner[, a Facebook policy lead,] focuses on defining policy for the site: guidelines about exactly what people should or shouldn't be allowed to do under what circumstances, and procedures for how violations are handled. These friendly and intelligent, young, blue jeans-wearing Californians *play the roles of lawmakers, judge, jury, and police all at the same time*. They operate a kind of private sovereignty in cyberspace.[63]

In this full excerpt, and in so much of her groundbreaking book, MacKinnon describes content moderation in the very words that Douek claims are absent from the standard picture.

Of course, any given work of scholarship argues and demonstrates much more than the single clause or quote to which it is reduced. But *Content Moderation as Systems Thinking* goes beyond reduction. This section takes the three substantive assertions in Douek's thesis in turn, comparing them with the text of the standard-picture scholarship and adding relevant citations from omitted scholarship.

1. The Standard Picture Sees Content Moderation Like a Real-World Government with Individual Adjudications, Bureaucracy, and Legislative Sessions. — In the standard scholarly picture of *Content Moderation as Systems Thinking*, "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

---

[62] *Id.* at 535 & n.22 (citing MACKINNON, *supra* note 24, at 153–54).

[63] MACKINNON, *supra* note 24, at 154 (emphases added).

judges.'"[64] The standard picture focuses overly on "individual posts" and assumes mistakenly that content moderation is only a "privatized hierarchical bureaucracy that applies legislative-style rules drafted by platform policymakers to individual cases and hears appeals from those decisions."[65]

Douek's use in quotes of "The New Governors" is a reference to my work of the same name, which I published in this *Review* in 2018.[66] In the early days of content moderation, many of the empirical intricacies of how and why private companies moderated speech on their platforms were a mystery.[67] Over three years, I interviewed former and current employees at large speech platforms, talked to members of civil society, and explored the existing literature.[68] A few things became clear: First, content moderation was nothing like the notice-and-takedown regime mandated by copyright law; instead, a much more complicated system was in place.[69] Second, though the substantive issues were distinct, the processes and systems that speech platforms were employing at scale were highly analogous to what my colleague Professor Rory Van Loo had characterized in the consumer law context in his article *The Corporation as Courthouse*.[70]

But Van Loo's comparison had limits, in part because content moderation wasn't about playing corporate middleman in buyer-seller contract disputes in the shadow of mandatory arbitration agreements. While particular commercial contexts certainly overlapped with speech platforms, the issues involved in content moderation had arguably higher stakes. As I described at the time, governing speech implicated human democratic participation, liberty, free expression, access to information, and community — but it also implicated child sexual abuse material, harassment, terrorism, fraud, hate speech, and misinformation.[71] Designing a system to deal with such trade-offs at a global scale in turn required infrastructure, processes, rules, people, and systems with complex motivations and influences.[72] I was far from the first or only person

---

64 Douek, *supra* note 1, at 529 (alteration in original).

65 *Id.* at 535.

66 *See* Klonick, *supra* note 4, at 1598.

67 *See* Catherine Buni & Soraya Chemaly, *The Secret Rules of the Internet: The Murky History of Moderation, And How It's Shaping the Future of Free Speech*, THE VERGE (Apr. 13, 2016, 10:30 AM), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech [https://perma.cc/D485-EL22]; Jeffrey Rosen, *Google's Gatekeepers*, N.Y. TIMES MAG. (Nov. 28, 2008), http://nyti.ms/20c9lqw [https://perma.cc/B8WH-WFF8].

68 *See* Klonick, *supra* note 4, at 1613–15, 1630–62, 1668–69. Works discussed include: CITRON, *supra* note 52; LAWRENCE LESSIG, CODE: VERSION 2.0 (2006); Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008); Balkin, *Digital Speech and Democratic Culture*, *supra* note 52; Balkin, *Old-School/New-School Speech Regulation*, *supra* note 52; Grimmelmann, *supra* note 24.

69 *See* Buni & Chemaly, *supra* note 67.

70 *See* Van Loo, *supra* note 26, at 559; Klonick, *supra* note 4, at 1647–48.

71 Klonick, *supra* note 4, at 1614, 1636, 1639, 1644 n.32, 1650, 1660, 1664.

72 *See generally id.*

to see it this way,[73] but my article added qualitative description and a theory of new governance at a moment where private control of public speech and its import to democracy became suddenly and massively visible.[74]

Over the course of seventy-three pages, *The New Governors* describes the multiple, and at times conflicting, motivations of private platforms to actively govern users' speech. It focuses on three of the largest user-generated content and speech platforms, all American companies, and describes the unique conditions of U.S. law that allowed for this self-regulation.[75] It details the dynamic system of ex ante automatic and ex post manual content moderation built over more than a decade.[76] It describes how the people, rules, and processes of that system are constantly changing in response to pluralistic systems of external influence from government, media, civil society, and individual users.[77] Its title and framing draw from Professor Jody Freeman's work, among others, in the "New Governance" movement that "proposes a conception of governance as a set of negotiated relationships between public and private actors."[78] It explicitly eschews First Amendment analogies and urges regulators to look at content moderation as a complex and iterative "system of governance."[79] None of this is included in Douek's summary of my article in a footnote parenthetical: "[D]escribing the three-tier structure of content moderation at Facebook."[80]

Other summations of the standard picture are also misleadingly reduced. Over the course of 214 pages, Tarleton Gillespie's 2018 book *Custodians of the Internet* describes social media platforms' content moderation as "functioning technical and institutional systems — sometimes fading into the background, sometimes becoming a vexing point of contention between users and platform."[81] But Douek samples only one line from the one chapter in which Gillespie describes just one part

---

[73] The work of Sarah Jeong, Adrian Chen, Catherine Buni, Soraya Chemaly, Adrian Lam, Cory Doctorow, and Daphne Keller, and Professors Jeffrey Rosen, Sarah T. Roberts, Jack Goldsmith, Tim Wu, David Post, Jack M. Balkin, and Lawrence Lessig, all flagged these difficult tradeoffs and their political stakes between 1995 and 2015.

[74] *The New Governors* was published in the *Harvard Law Review* on April 10, 2018, the same day Facebook CEO Mark Zuckerberg first testified before the U.S. Senate over the platform's role in election integrity. *Compare* Klonick, *supra* note 4, at 1667–68 (published online Apr. 10, 2018), *with* Camila Domonoske, *Mark Zuckerberg Tells Senate: Election Security is an "Arms Race"*, NPR (Apr. 10, 2018, 2:30 PM), https://www.npr.org/sections/thetwo-way/2018/04/10/599808766/i-m-responsible-for-what-happens-at-facebook-mark-zuckerberg-will-tell-senate [https://perma.cc/4PB8-2M3U].

[75] Klonick, *supra* note 4, at 1603–13.

[76] *Id.* at 1635–39.

[77] *Id.* at 1648–58.

[78] Jody Freeman, *The Private Role in Public Governance*, 75 N.Y.U. L. REV. 543, 543 (2000). Douek also cites to Freeman and adopts this same framing. Douek, *supra* note 1, at 530 n.10.

[79] Klonick, *supra* note 4, at 1669.

[80] Douek, *supra* note 1, at 535 n.22.

[81] GILLESPIE, *supra* note 4, at 6.

of content moderation as emblematic of the dominant standard picture overly focused on individual ex post manual content decisions.[82]

Content Moderation as Systems Thinking repeatedly mischaracterizes the scholarship's empirical observations as normative arguments. It was not Professor David Kaye, for example, who characterized Facebook's policy process as a "mini legislative session" but a Facebook employee.[83] Douek also uses a quote from Professor Kyle Langvardt's article *Can the First Amendment Scale?* as evidence of the standard picture viewpoint:

> 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.[84]

Langvardt himself is not *advocating* for this approach, but merely stating that such an approach exists. Indeed, in the very same passage he expressly acknowledges that "those [ex post] tools also have their limits,"[85] largely because individual challenges to removal decisions will not "translate to anything systemic."[86] Langvardt's point is exactly the opposite for which he is cited and in fact, makes the very same argument that Douek claims as part of her novel thesis.

2. The Standard Picture Misses the Trade-Offs, Outside Influence, and Automatic Side of Content Moderation. — Content Moderation as Systems Thinking argues that the standard-picture scholarship ignores "the wide variety of *ex ante trade-offs* that content moderation institutional designers have to engage with."[87] It does not understand that "content moderation bureaucracies are a 'they' not an 'it':" composed of a "*wide diversity of institutions involved in content moderation outside the hierarchical bureaucracy* that is the content moderation appeals system."[88]

---

[82] Douek, *supra* note 1, at 535 n.22 ("[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." (alteration in original) (quoting GILLESPIE, *supra* note 4, at 116)).

[83] *Compare id.* at 535 n.22 (describing Kaye as categorizing a policy meeting at Facebook as a "mini-legislative session" where questions of notice, due process, and appeal are "exactly the right questions you would hope Facebook would be asking itself"), *with* KAYE, *supra* note 26, at 53–58 (describing a *Facebook employee* calling the session the author attended a "mini-legislative session," *id.* at 54, but calling the session "a lengthy *process*, one that has involved internal review and the solicitation of views from outside the company," *id.* at 57–58 (emphasis added), and concludes when the "group finds consensus on the change and the responsible team members . . . move toward implementing the new policy," *id.* at 58).

[84] Douek, *supra* note 1, at 535 n.22 (quoting Langvardt, *supra* note 44, at 298).

[85] Langvardt, *supra* note 44, at 298.

[86] *Id.* at 299.

[87] Douek, *supra* note 1, at 539 (emphasis added).

[88] *Id.* (emphasis added).

Automatic and ex ante content moderation have always been part of the scholarly content moderation conversation. *New Governors* reorganized and restructured a taxonomy created by Professor James Grimmelmann in his formative work *The Virtues of Moderation*.[89] Grimmelmann's piece was published in 2015 in the *Yale Journal of Law and Technology*.[90] Douek cites to it in her second footnote,[91] but somehow it is not part of the standard picture[92] despite the fact that Grimmelmann describes moderation in the following terms: "[M]oderation can be carried out manually, by human moderators making individualized decisions in specific cases, or automatically, by algorithms making uniform decisions in every case matching a specified pattern."[93]

Grimmelmann describes much of this automatic moderation as "ex ante" because it happens before publication.[94] In updating Grimmelmann's taxonomy in *New Governors*, I added an important description: "*The vast majority*" of content moderation, I wrote in 2018, "is an automatic process run largely through algorithmic screening without the active use of human decisionmaking."[95]

This was an important distinction to make because at the time, and still today, people were largely unaware of two huge parts of their online lives: one, that content moderation was happening at all; and two, that if it was happening, humans were involved. To the former, the fact that ex ante automatic content moderation stopped content from *ever* appearing on another user's Facebook feed had different implications for speech (think prior restraint) and the system of speech governance than ex post reactive manual content moderation had.[96] The adjective "reactive" in this description spoke to the platform reacting to users flagging problematic "ex post" (published) content, while "manual" referred to the human content moderator who would then look at the flagged content and decide whether to remove it from the site.[97] In her book *Behind the Screen: Content Moderation in the Shadows of Social Media*, Professor Sarah T. Roberts describes the moment when she first discovered from a 2010 *New York Times* article that humans were doing this review:

> I forwarded the article to a number of friends, colleagues and professors, all longtime internet users like me, and digital media and internet scholars themselves. "Have you heard of this job?" I asked. "Do you know anything

---

89 *See* Klonick, *supra* note 4, at 1635–38, 1635 n.261.
90 Grimmelmann, *supra* note 24, at 42.
91 Douek, *supra* note 1, at 528 n.2.
92 *Id.* at 535 n.22.
93 Grimmelmann, *supra* note 24, at 55.
94 *Id.* at 67.
95 Klonick, *supra* note 4, at 1636 (emphasis added).
96 *See id.* at 1636–38.
97 *Id.* at 1635 (emphasis omitted).

> about this kind of work?" None of them had . . . . They, too, were transfixed.[98]

Even eight years after the *New York Times* article and Roberts's revelation, there was relatively little awareness about how content moderation worked or that there were humans in the loop. Many individuals simply thought that "computers" adjudicated content,[99] somehow able to grok, for example, the invisible element of user intent that makes a picture of a topless woman posted as protest different from a picture of a topless woman posted as pornography. *New Governors'* description of ex ante automatic content moderation focused on the proliferation of the use of "hashing" to check a known universe of banned content against something that is uploaded, rather than the use of photo recognition or natural language bans.[100] But it also described "ex post reactive manual" content moderation — humans posting, humans flagging those posts, and humans reviewing for violations — and how that system iterated on itself over time but also sent signals back to the ex ante system so that that automatic process regularly changed.[101]

Whether automatic or human, content moderation considerations necessarily required "trade-offs" — between how proactive a platform was in removing content, how to select the revisions of "standards to rules" it enforced,[102] and how much it relied on "automatic ex ante"[103] versus "ex post reactive manual"[104] content moderation done by individuals.[105] Perhaps the best and most recent description of these tradeoffs in ex ante content moderation comes in Professor Hannah Bloch-Wehba's article, *Automation in Moderation,* published in the *Cornell International Law Journal* in 2020.[106] It is worth noting that in a few years since publication, the paper has been widely read,[107] yet it is not included as part of the "standard picture" of content moderation. Bloch-Wehba surveys the scholarly history of automatic content moderation and describes the current state of technology. Her normative takeaway is powerful and clear: "[N]ew automation techniques exacerbate existing risks to free speech and user privacy, and create new sources of information . . . for surveillance, raising concerns about free association, religious freedom, and racial profiling . . . [and] worsens transparency and accountability deficits."[108]

---

98 ROBERTS, *supra* note 52, at 22.
99 *See generally id.*
100 Klonick, *supra* note 4, at 1637.
101 *Id.* at 1637–38.
102 *Id.* at 1632.
103 *Id.* at 1637.
104 *Id.* at 1638.
105 *Id.* at 1668.
106 Bloch-Wehba, *supra* note 24.
107 As of April 2023, the article has 763 downloads on SSRN. *See Automation in Moderation*, SSRN (Feb. 11, 2021), https://papers.ssrn.com/abstract=3521619 [https://perma.cc/2AYV-YY5F].
108 Bloch-Wehba, *supra* note 24, at 43.

Grimmelmann's, Roberts's, Bloch-Wehba's, and my own work are not alone in describing content moderation not just as individual posts but also as a complex mix of both ex ante and ex post content adjudication involving difficult tradeoffs. For example, Gillespie's *Custodians of the Internet* describes the processes of both ex post and ex ante moderation throughout the work.[109] So too does Kaye, at the outset of *Speech Police*:

> The enormous volume of uploaded content requires that the company rely on two tools to surface potentially problematic or illegal content: humans who comb through and report content, and algorithmic automation, or Artificial Intelligence. Ideally, flagged content would undergo human evaluation before it is taken down, whether it results from human or algorithmic flagging. But that's not always the case. Both human and algorithmic flagging can lead to mistaken deletions or blockings, or ones that activists or governments may simply disagree with.[110]

In 2019, a terrorist live streamed the shooting of a mosque in Christchurch, New Zealand, on Facebook.[111] Though the initial video post was removed relatively quickly from the platform, it had been captured by trolls on notorious sites like 8chan.[112] Despite Facebook having added a hash for the video to its automatic database so it couldn't be reposted, for days after the tragedy, trolls uploaded copies of the live steam manipulated to get past automated ex ante detection and reappear on the platform.[113] In a piece for the *New Yorker* following the attack, I described the global teams of individuals that worked around the clock to chase and take down the video — and ultimately devise a new system of hashing and automated behavioral identification that couldn't be manipulated by such trolls.[114]

Though not included in *Content Moderation as Systems Thinking*, there is a plethora of scholarship that specifically discusses how the rules applied by these automatic or manual processes are created, changed, or eliminated through a global *system* of engineers, policymakers, activists, platform managers, and many others. I documented this "pluralistic system of influence"[115] from government, media, civil society, and individuals, but especially the influence of government and platform cooperation, in section III.C of *New Governors*.[116] This was also a central thesis of Professors Jack Goldsmith and Tim Wu's early book, *Who Controls the Internet?*, which predicted, accurately, how governments

---

109 *See generally* GILLESPIE, *supra* note 4.

110 KAYE, *supra* note 26, at 26.

111 Kate Klonick, *Inside the Team at Facebook that Dealt with the Christchurch Shooting*, NEW YORKER (Apr. 25, 2019), https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting [https://perma.cc/TC4Z-PPHP].

112 *Id.*

113 *Id.*

114 *Id.*

115 Klonick, *supra* note 4, at 1648.

116 *See id.* at 1648–58.

would come to exercise geopolitical power through control and lobbying of internet stakeholders.[117] MacKinnon's *Consent of the Networked* is almost entirely devoted to the development of this online balance of power between governments and platforms — as the book's tacit reference to John Locke's formulation canonized in the United States Declaration of Independence suggests.[118] MacKinnon also spends much of her time on the development of multistakeholder solutions to these problems, facilitated by international law.[119] This is a theme reexamined by Kaye's *Speech Police*, which updates MacKinnon's formulations with modern examples from the international human rights perspective.[120]

Delegated decisionmaking is also discussed by the standard picture scholarship. Informal relationships between third-party experts and platforms is talked about in *New Governors*,[121] while I documented the set up and influence of Facebook's Oversight Board in 2020 in the *Yale Law Journal*.[122] Most notably, in her recent essay for the *Harvard Law Review Forum*, *Facebook's Faces*, Chinmayi Arun adeptly discusses the complicated and dynamic relationship between individuals inside the platforms and individuals outside.[123] "Facebook engages with states and publics through multiple parallel regulatory conversations, further complicated by the fact that Facebook itself is not a monolith," Arun writes.[124] "Facebook has many faces — different teams working towards different goals, and engaging with different ministries, institutions, scholars, and civil society organizations. It is also internally complicated, with staff whose sympathies and powers vary and can be at odds with each other. Content moderation takes place within this ecosystem."[125]

3. The Standard Picture Is Preoccupied with First Amendment Analogy. — Finally, *Content Moderation as Systems Thinking* argues that the standard picture "assumes the necessity of a model of speech governance and the judicial role adapted from the First Amendment context"[126] and is "pervaded by First Amendment analogies."[127] While content moderation scholarship certainly argues that First Amendment principles have had an implicit and normative role in shaping content

---

117 *See generally* GOLDSMITH & WU, *supra* note 29.
118 MACKINNON, *supra* note 24, at 164.
119 *Id.* at 200–40.
120 *See* KAYE, *supra* note 26, at 51.
121 Klonick, *supra* note 4, at 1655–57.
122 Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L.J. 2418, 2425 (2020).
123 Arun, *supra* note 27, at 236.
124 *Id.*
125 *Id.*
126 Douek, *supra* note 1, at 536.
127 *Id.* at 556.

moderation systems, it is inaccurate to describe it as frequently dominated by First Amendment analogies.

Douek's primary citation for this claim is to *New Governors*, but the relevant text from the pages she cites states exactly the *opposite* of her assertion. From *New Governors*:

> *Though they might not have "directly imported First Amendment doctrine,"* the normative background in free speech had a direct impact on how they structured their policies. Wong, Hoffman, and Willner all described being acutely aware of their predisposition to American democratic culture, which put a large emphasis on free speech and American cultural norms. Simultaneously, there were complicated implications in trying to implement those American democratic cultural norms within a global company.[128]

This is not the only point at which I explicitly eschew First Amendment analogy as the standard for understanding private content moderation: I do so five times throughout the article, including in the abstract and introduction. The following excerpts are all from *New Governors*:

- "This Article argues that to best understand online speech, *we must abandon traditional doctrinal and regulatory analogies* and understand these private content platforms as systems of governance."[129]
- "[T]his Article argues that analogy purely under First Amendment doctrine should be largely abandoned."[130]
- "The law reasons by analogy, yet none of these analogies to private moderation of the public right of speech seem to precisely meet the descriptive nature of what online platforms are, or the normative results of what we want them to be."[131]
- "Thinking of online platforms from within the categories already established in First Amendment jurisprudence — as company towns, broadcasters, or editors — misses much of what is actually happening in these private spaces."[132]

Nor am I alone in repeatedly and categorically denying the applicability of First Amendment analogies to online speech governance, though I am the only one cited by Douek. In *Speech Police*, Kaye disavows the views of "American legislators and policymakers [who] . . . are constitutionally myopic in their rigid understanding and politicization of First Amendment values."[133] Outside the standard picture, Professor Jack Balkin writes in *Free Speech Is a Triangle,* published in the *Columbia Law Review* in 2019, that "the best alternative to

---

128 Klonick, *supra* note 4, at 1621 (emphasis added).
129 *Id.* at 1599 (emphasis added).
130 *Id.* at 1602–03.
131 *Id.* at 1662.
132 *Id.*
133 KAYE, *supra* note 26, at 17.

this autocracy *is not the imposition of First Amendment doctrines by analogy to the public forum or the company town*."[134]

C. Blaming the Standard Picture

The final part of *Content Moderation as Systems Thinking*'s central claim is that the standard scholarly picture "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."[135]

Even if a cohesive standard picture of content moderation scholarship exists, *Content Moderation as Systems Thinking* never offers any evidence that it is *the scholarship* that has led to lawmakers' incomplete understanding of online speech or flawed regulatory proposals. Indeed, the very words describing the standard picture's focus on "paradigm cases"[136] as "*high-profile* content moderation controversies"[137] that "*dominate media headlines*"[138] suggest that such emphasis is due to *the media's construction*, not scholarship's.[139] It seems nonsensical to suggest a small cohort of interdisciplinary academics are to blame for lawmakers' obsession with individual speech cases, rather than the press or lawmakers themselves.[140] Arguing that *the media* has presented an

---

[134] Balkin, *supra* note 23, at 2025 (emphasis added). Douek cites to *Free Speech Is a Triangle* in support of her thesis, Douek, *supra* note 1, at 556 n.144, and does not include any of Balkin's work in the standard picture, though roughly one-third of Balkin's article favorably adopts the description of *New Governors* and then builds on its premises, *see* Balkin, *supra* note 23, at 2018 n.24.

[135] Douek, *supra* note 1, at 531.

[136] *Id.*

[137] *Id.* at 536 (emphasis added).

[138] *Id.* at 529 (emphasis added).

[139] *See id.*

[140] Douek also errs in her characterization of current regulation as myopically focusing on "individual procedural rights" in the first place, Douek, *supra* note 1, at 531, particularly in her reliance on the European Commission's Digital Services Act (DSA), 2022 O.J. (L 277) 1. Douek suggests that the DSA does not account for different types or sizes of platforms. Douek, *supra* note 1, at 567. But the DSA is specifically drawn around different types and sizes of platforms and changes their legal duties accordingly. *See, e.g.*, *The Digital Services Act: Ensuring a Safe and Accountable Online Environment*, EUR. COMM'N, https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act-ensuring-safe-and-accountable-online-environment_en [https://perma.cc/73HN-EQ36] (explaining to the public the distinctions between intermediary services, hosting services, online platforms, and very large online platforms under the DSA). The DSA, she also states, "ignores the institutional diversity and ex ante design choices." Douek, *supra* note 1, at 565. But this is also not true. The DSA includes multiple provisions that address automatic and ex ante content choices. *See, e.g.*, Bengi Zeybek, *The DSA and the Risk-Based Approach to Content Regulation: Are We Being Pulled into More Advanced Automation?*, DSA OBSERVATORY (Oct. 1, 2021), https://dsa-observatory.eu/2021/10/01/the-dsa-and-the-risk-based-approach-to-content-regulation-are-we-being-pulled-into-more-advanced-automation [https://perma.cc/H2TV-MB92] (discussing "the role of automated tools with regard to compliance with the risk assessment and mitigation obligations under Articles 26 and 27" and "the implications of the risk-based approach for automated content moderation tools"). Finally, Douek entirely omits the existence of the Digital Markets Act (DMA), 2022 O.J. (L 265) 1, the companion

oversimplified version of online speech and content moderation would have been a far more accurate, albeit narrower, claim. Douek perhaps realizes this: despite leveling the blame solely at scholars, only roughly half of the citations in her footnote describing the standard picture are to academic sources, and the remainder are reports or media coverage.[141]

II. WHY IT MATTERS

Despite my criticism of *Content Moderation as Systems Thinking*, I do not at all disagree with the overall theory it proposes. Nor do I take issue with the idea that content moderation should be seen systemically, focused on "wholes and interrelationships rather than parts."[142] I agree that content moderation platforms are much more than post-by-post decisionmaking but instead complex and dynamic systems. And I agree that offline models of adjudication and the First Amendment provide a poor framework for understanding how online speech platforms work. I agree that content moderation — truly, all line drawing around speech — is full of tradeoffs and that perfection is impossible. Indeed, it would be hypocritical of me not to agree, because I and so many of the people I admire in this field have said so much of this before. But ultimately, the main reason that I contest the construction and characterization of a "standard picture" of content moderation is that it risks serving as a misleading premise to a shortsighted set of reforms.

A. Government-Mandated Transparency and Process Cannot Solve the Problem of Transparency and Process Theater

The central harm of the standard picture and its influence, Douek claims, is that content moderation reform has overly focused on transparency reports, individual procedural rights, and individual content appeals.[143] Though the blame is misplaced, the critique is valid. Individual content decisions are imperfect mechanisms for signaling representative change back to the system, and at scale they are often inadequate remedies for users, coming too late and offering too little. The result is not meaningful changes and accountability, she argues, but simply the performance of accountability — "process theater"[144] or in the case of transparency reports, "transparency theater."[145]

---

competition legislation to the DSA. The DMA does much of what Douek claims regulators are not doing — it enforces with teeth. Specifically, the DMA lays enormous financial penalties for platforms that are out of compliance with its regulatory requirements like interoperability and user access to data, *see id.* — some of the very concepts Douek claims current regulatory proposals are missing.

141 Douek, *supra* note 1, at 535 n.22.

142 *Id.* at 530.

143 *Id.* at 565–77.

144 *Id.* at 577.

145 *Id.* at 572; *see also id.* at 572–82.

An example of the worst of these performances is the Facebook Oversight Board (FOB), the independent adjudicator set up by Meta in 2020 to hear content appeals and issue decisions. "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," Douek writes.[146] This approach, she claims, is full of "futility and failures"[147] that miss aggregate harms, broken AI, and operational mistakes.[148]

This might well be true, but it is hard to understand how emphasizing individual rights has caused this to be the case, or how dismantling such processes will solve it. Moreover, it would seem that a systemic solution like the one Douek proposes would take *both* into account, allowing the Oversight Board to be a dynamic solution for content moderation reform, not a panacea. I have said as much in my prior writing — and, somewhat confusingly, so has Douek. The FOB "will not solve all our problems with social media," she acknowledged in 2020, listing the problems the Board cannot address such as AI bias and independent researcher access.[149] But despite these shortcomings, she argues, the Board has an important role to play:

> Currently, some of the most consequential decisions about the way information flows through society occur behind closed doors with minimal public justification and in a way that is influenced by business imperatives. This is at odds with how essentially every jurisdiction with free speech traditionally thinks about it, which is that any restrictions on speech should be specified clearly in advance, applied consistently, and subject to careful scrutiny. This is the check that the FOB can bring to Facebook's content moderation ecosystem.[150]

It is hard to square this acknowledgement and countless other writings by Douek that praise the Board for bringing some amount of transparency and process to content moderation with the final section of *Content Moderation as Systems Thinking*. Nor is it clear how the reforms that the article proposes instead would escape these problems of "performance" or "theater." The modest proposals in the final Part of the article include structural and procedural requirements like annual content moderation plans and compliance reports, quality assurance, and audits to be performed by government agencies.[151] I will return to the solution of government agencies as enforcement in a moment, but as a foundational matter simply adding more transparency is not a solution

---

146 *Id.* at 568.

147 *Id.*

148 *Id.* at 568–70.

149 Evelyn Douek, *"What Kind of Oversight Board Have You Given Us?"*, U. CHI. L. REV. ONLINE (May 11, 2020), https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek [https://perma.cc/Z8EN-6TUP].

150 *Id.*

151 Douek, *supra* note 1, at 584–606.

to performative transparency or theater. Indeed, the opposite has been proven true. This is the "transparency paradox" as Professor Ethan Bernstein coined and empirically demonstrated, in which increasing the size and salience of an audience paradoxically reduces sincerity and heightens performance. "Analogously, increasing observability in a factory may in fact reduce transparency, which is displaced by illusory transparency and a myth of learning and control, by triggering increasingly hard-to-detect hiding behavior," writes Bernstein.[152] This does not mean that there is no value in transparency, or that such attempts should be abandoned, but it does mean that many of Douek's suggested reforms might indeed only serve to heighten the very process and transparency "theater" she critiques, rather than resolve them.

B. Speech Is Legally and Phenomenologically Special — And It Should Be

Critiquing, even condemning, past reform efforts would not be problematic if they were not presented in false dichotomy with Douek's own solutions for reform and if some of her reforms weren't so potentially dangerous to democracy. Almost all the proposals in Part IV of *Content Moderation as Systems Thinking* have been previously proposed and are as modest as she suggests. The unique element among those reforms is that they be enforced by a government administrative agency.

To square this prescription of an administrative agency for content moderation with the First Amendment, Douek argues that "[s]peech [i]s [n]ot [s]o [s]pecial."[153] This framing — "must speech be special?" — is borrowed from Professor Frederick Schauer's work of that name, but critically it neglects to mention that Schauer's titular question is not rhetorical. (Indeed, after a formal logic analysis, he concludes the opposite of what Douek suggests: yes, speech must and should be special.[154]) Instead, her argument centers on the idea that speech need not be special because it also can be commercial in nature. "[M]any canonical content moderation controversies are about commercial interests,"[155] she writes, referencing controversies around Nazi memorabilia sold on Yahoo!, or eBay delisting Dr. Seuss books,[156] "but they get 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*. But . . . speech!"[157]

---

[152] Ethan S. Bernstein, *The Transparency Paradox: A Role for Privacy in Organizational Learning and Operational Control*, 57 ADMIN. SCI. Q. 181, 216 (2012).

[153] Douek, *supra* note 1, at 556.

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

[155] Douek, *supra* note 1, at 558.

[156] *Id.* at 558 n.162.

[157] *Id.* at 558–59 (ellipsis in original).

But that commercial interests also exist alongside speech interests in content moderation hardly seems wholly damning of the unique place for speech in the law and democracy, generally. Books and newspapers are sold and published by profit-driven entities, and not only are they considered speech, the institutions that produce them have their own First Amendment protections.[158] Nor is the framing as speech cases necessarily an indictment of speech as a special category, so much as skilled lawyering. Indeed, Douek's complaint seems to be more with uneducated journalists and Americans than with the law itself: "On the current state of the law, there is not even a colorable First Amendment claim against platforms for restricting users' speech. Yet cries of 'First Amendment!' or 'Free Speech!' abound when they do."[159] That many are not aware that the First Amendment only applies to *government* restriction of speech and not *any* restriction of speech seems more a failure of civic education, than an indictment of a near-universally agreed-upon human right.

It is not entirely clear why Douek bothers arguing that speech is not special until you understand that this in some sense a paper *about administrative law* that is arguing *for administrative law solutions*. Any chance at such heavy-handed government regulatory reform over private speech rights in content moderation necessitates arguing that perhaps speech isn't so special and the First Amendment shouldn't prevent such a reform. Without arguing that speech isn't special, the administrative agency solution of Part IV of *Content Moderation as Systems Thinking* is even less feasible than it otherwise would be.

Finally, it is paradoxical after an entire paper lamenting the impossible tradeoffs and arguing that perfection is impossible in content moderation that Douek argues for administrative agency oversight not just to police the reforms she proposes, but to assure content moderation "quality."[160] Douek admits that the idea of quality is a "deeply contested concept" and briefly lists several diverse factors that could possibly measure quality in content moderation.[161] Her article ends with the assertion that "[t]he only thing worse than trying to define 'quality' is *not* trying."[162]

I am not so sure. It would seem like one of the worst things you could do for democracy is to give a government agency blank-check authority to enforce an undefined standard like "quality" over its citizens' speech. This is all the truer when such government control would include "creating more specific standards and mandates in the future"

---

158 *See, e.g.*, N.Y. Times Co. v. Sullivan, 376 U.S. 254, 270 (1964).
159 Douek, *supra* note 1, at 558.
160 *Id.* at 584.
161 *Id.* at 601.
162 *Id.*

for ex ante content moderation — the most invisible, and therefore potentially censorial, area of speech governance.[163]

Accuracy in representing scholarship and thinking through the consequences of potential reform matters, because solving the problem of online content moderation is not an academic question and it is not a game. It is a very real problem with real-world consequences across almost every dimension of global society. Changes in U.S. law or policy around online speech will have dramatic effect far outside the United States's borders. Speech, particularly speech published in this law review, can have great impact and significance on the world.

CONCLUSION

It is not enough to just generally describe some things as "systemic," and others as not, to take a systems-thinking approach. Systems thinking is a dynamic and powerful tool of description to understand complex phenomena. There is no wonder it is perhaps best known in understanding biological ecosystems. It is both the ocean *and* the wave, the fish below *and* the boat above.[164]

To best understand content moderation as systems thinking, one would have to accede that content moderation contains individual decisions, automations, governance, governments, external influence, internal politics, constitutions, norms, legality, human judgment and biases, administration, bureaucracy, multistep processes, long legislative-like meetings, people, corporate courthouses, actual courthouses, stakeholders, economies, the media, and iterative dynamic changes. To understand content moderation as systems thinking, one would have to rely on the long history of scholarship that lays each of these elements and systems out. And in doing so, one would have to acknowledge the vastness of the ocean and the insignificance of a single wave.

---

[163] *Id.* at 586.

[164] Arnold & Wade, *supra* note 17, at 671.