SEPARATION OF FUNCTIONS FOR AI: RESTRAINING SPEECH REGULATION BY ONLINE PLATFORMS

by
Niva Elkin-Koren[*] *& Maayan Perel*[**]

The Free Speech Clause of the First Amendment of the U.S. Constitution restricts government regulation of private speech. However, it generally does not apply to private management of speech. New forms of speech regulation by online platforms disrupt this constitutional framework. Platforms, such as Google, Facebook, and Twitter, are responsible for mediating much of the public discourse and governing access to speech and speakers around the world. These private businesses match users and content in whatever way best benefits their commercial interests. At the same time, however, they exercise regulatory power when they filter, block, and remove content at the request of governmental agents or state actors. Consequently, platforms effectively blend law enforcement and adjudication powers, and sometimes even lawmaking powers.

Courts and scholars who tackle speech regulation by platforms have basically relied on the well-settled constitutional divide between private functions and governmental ones. To the extent that platforms exercise governmental powers in allowing or banning speech or speakers, platforms should be subject, as the argument goes, to public law principles of accountability, legitimacy, oversight, and power separation.

In this paper, we question this approach. As a practical matter, the public/private framework presumes that public functions of a private entity could be neatly separated from its standard business affairs. We argue that with the increasing use of Artificial Intelligence (AI) by platforms for content moderation, the public, law enforcement functions are integrated with the private, business functions that are driven by commercial interests. The same technical design which is used for targeted advertising and for curating personalized content is also deployed for monitoring and censoring online content. Using machine learning, the system is informed by the same labeling of users and

[*] Prof. Niva Elkin-Koren is a Professor at the University of Haifa Faculty of Law and a Faculty Associate at the Berkman Klein Center for Internet & Society at Harvard University.

[**] Dr. Maayan Perel is an Associate Professor at Netanya Academic College, Faculty of Law, and a Senior Research Fellow at the Haifa Center for Law & Technology, University of Haifa Faculty of Law. This Research was supported by the Israel Science Foundation (grant No. 1820/17). We thank Ellen Goodman, Amélie Heldt, Yifat Nahmias, Moran Yemini, and the participants of the 2019 international conference on Harmful Online Activity and Private Law held at the Hebrew University as well as the participants of the 2019 Annual Israeli Intellectual Property Scholars Conference held at Tel Aviv University for helpful comments.

content, and makes use of the same application programming interfaces (API), learning patterns, and software. Consequently, decisions on removal of speech, for (public) law enforcement purposes, are driven by the same data, algorithms, and optimization logic, which are also underlying all other functions performed by the platform. Therefore, the use of AI in content moderation calls for a fresh approach to restraining the power of platforms and securing fundamental freedoms in this environment.

This paper takes a design perspective to speech regulation. It contends that the normative distinctions between public and private functions could be upheld in online content moderation, provided that these distinctions are embedded in the system design. It introduces "separation of functions," a novel approach to restraining the power of platforms while enhancing the accountability in AI-driven content moderation systems. We propose to facilitate independent tools embedding public policy. These tools would run on the platforms' data and would include their own optimization processes informed by public policy. Such separation between independent public tools and private data may enhance public scrutiny of law enforcement speech restrictions, which are a traditionally exclusive public function. This functional separation may also facilitate competition among different players who may enrich the design of speech regulation and mitigate biases. Finally, we explore the implications of this approach and discuss its possible limitations.

I. INTRODUCTION

In the United States, the First Amendment protects *private* speech from "the most coercive technique of the *government*—direct and coercive punishment of

disfavored speakers."[1] The assumption is that the "marketplace of ideas," where all players can freely interact without government intervention, will secure democratic self-governance and civil liberties.[2] However, today's online speech environment proves otherwise: even though it is *private* platforms who practically govern "new-school" speech regulation,[3] the marketplace of ideas seems to malfunction: a disturbing spread of unwanted speech accompanies occasional restrictions of desired speech.[4] In many respects, governments deputize private platforms as censors, and practically control billions of online speakers around the world and shape the public discourse while bypassing constitutional constraints.[5] This relationship between the state and private sectors has been called the "invisible handshake,"[6] and "collateral censorship,"[7] or "censorship by proxy."[8] Could speech regulation by platforms be bound by the First Amendment?

Courts and scholars who address speech regulation by platforms rely on a well-settled legal principle in constitutional scrutiny: the distinction between private functions and governmental ones. The non-delegation doctrine, for instance, sets limits on congressional power to delegate its legislative power directly to the *private* sector.[9] The rationale is that private actors are not sufficiently bound by constitutional principles of accountability, transparency, and legitimacy,[10] and may not adequately represent the public interest, as they might be biased towards their

---

1 Tim Wu, *Is the First Amendment Obsolete?*, 117 MICH. L. REV. 547, 568 (2018) (emphasis added).

2 United States v. Rumely, 345 U.S. 41, 56 (1953) (Douglas, J., concurring).

3 Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2308–10 (2014); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1599–1602 (2018).

4 Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1355–56 (2018).

5 *See* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, 29 n.9 (Hoover Institution, Aegis Series Paper No. 1902, 2019), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf (referring to different "scholars [who] have long anticipated the emergence of internet companies as 'private surrogates' allowing governments to bypass 'pesky constitutional constraints.'") (quoting James Boyle, *A Nondelegation Doctrine for the Digital Age?*, 50 Duke L.J. 5, 10–11 (2000) [hereinafter Boyle, *Nondelegation Doctrine*]; James Boyle, *Foucault in Cyberspace: Surveillance, Sovereignty, and Hardwired Censors*, 66 U. CIN. L. REV. 177, 201 (1997) [hereinafter Boyle, *Foucalt in Cyberspace*] (describing information "guarded by digital fences which themselves are backed by a state power maintained through private systems of surveillance and control").

6 Michael D. Birnhack & Niva Elkin-Koren, *The Invisible Handshake: The Reemergence of the State in the Digital Environment*, 8 VA. J.L. & TECH. 6, ¶ 2 (2003); Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105, 107 (2016) (also using the "invisible handshake" terminology).

7 Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 16–17 (2006).

8 Michael I. Meyerson, *Authors, Editors, and Uncommon Carriers: Identifying the "Speaker" Within the New Media*, 71 NOTRE DAME L. REV. 79, 118 (1995).

9 Kimberly N. Brown, *Public Laws and Private Lawmakers*, 93 WASH. U. L. REV. 615, 660 (2016) (arguing that the nondelegation doctrine warrants constitutional scrutiny of executive branch outsourcing of legislative power to private parties).

10 *Id.* at 618, 620–21.

commercial interests.[11] Similarly, under the "state action" doctrine,[12] the Constitution generally applies only to government conduct and does not prohibit the deprivation of constitutional rights by private actors.[13] The separation of powers is another important aspect of the public/private divide, which ensures adequate checks and balances in the exercise of governmental power to facilitate oversight and safeguard against abuse of power.[14]

These fundamental principles of constitutional law assume that it is possible to distinguish between governmental actions and conduct by private actors, even if the applications of this divide to different regulatory regimes may differ.[15] As a practical matter, the public/private framework further presumes that public functions of a private entity could be neatly separated from its standard business affairs. Specifically, in the field of communications, the Supreme Court recently held in *Manhattan Community Access Corp. v. Halleck* that the Free Speech Clause of the First Amendment of the United States Constitution prohibits only governmental, not private, abridgment of speech.[16] At the same time, however, private actors might be held liable for violating the First Amendment under the state action doctrine when they act on behalf of the government or perform a function that is normally done by the government.[17] While the operation of a public forum for speech is not bound by governmental constraints on speech, a private entity which is performing "a traditional, exclusive public function" would be.[18]

The constitutional divide between public and private also applies to content moderation by platforms. Platforms are private businesses which are matching users and content in the ways that best benefit their commercial interests.[19] At the same time, platforms, such as Google, Facebook, and Twitter, are responsible for mediating much of the public discourse and governing access to speech and speakers around the world. Consequently, they have become ideal partners for governments in performing civil and criminal law enforcement.[20]

---

[11] The Supreme Court in *Carter v. Carter Coal Co.* has warned against the risks of delegating powers to a private party "whose interests may be and often are adverse to the interests of others in the same business." Carter v. Carter Coal Co., 298 U.S. 238, 311 (1936) (striking down a statute authorizing local coal boards to determine coal prices and employee wages and hours, based on the Commerce and Due Process Clauses).

[12] Lillian BeVier & John Harrison, *The State Action Principle and Its Critics*, 96 VA. L. REV. 1767, 1769 (2010).

[13] The Civil Rights Cases are usually credited with being the origin of the state action requirement. *See* Civil Rights Cases, 109 U.S. 3, 11 (1883); *see also* BeVier & Harrison, *supra* note 12, at 1769; Erwin Chemerinsky, *Rethinking State Action*, 80 NW. U. L. REV. 503, 505, 507 (1985).

[14] *See infra* Part II.C.

[15] Peter Cane, *Public Law and Private Law: The Study of the Analysis and Use of a Legal Concept*, *in* OXFORD ESSAYS IN JURISPRUDENCE 57, 57–61 (John Eekelaar & John Bell eds., 1987).

[16] Manhattan Cmty. Access Corp. v. Halleck, No. 17-1702, slip op. at 2 (587 U.S. __ June 17, 2019).

[17] *Id.* at 6.

[18] *Id.* at 4.

[19] *See infra* Part III.D.

[20] Annemarie Bridy, *Graduated Response and the Turn to Private Ordering in Online Copyright Enforcement*, 89 OR. L. REV. 81, 83–84 (2010); Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J.L. & TECH. 253, 255–57 (2006).

Scholars have shown how platforms "can take on and displace traditional state functions, operating the modern equivalent of the public square or the post office, without assuming state responsibilities."[21] Platforms exercise regulatory power when they filter, block, and remove content, at the request of governmental agents or state actors, defining the practical benchmark for illegality, adapting it to the changing circumstances, and applying it to particular expressions.[22] As a result, platforms effectively blend law enforcement and adjudication powers, and sometimes even lawmaking powers, and yet are not subjected to adequate constitutional checks.[23]

The ongoing 2020 Covid-19 pandemic has further demonstrated that the boundaries between public and private in speech regulation are blurred. In the wake of the crisis there was a growing pressure on platforms to act against the proliferating of misinformation and conspiracy theories which were threatening to put public health and safety at risk.[24] The major platforms, including Facebook, Google and Twitter, announced that they would ban and take down conspiracy theories and misleading false claims regarding the health crisis.[25] As reported extensively, "platforms are proudly collaborating with one another, and following government guidance, to censor harmful information related to the coronavirus."[26] Yet, the health crisis was subsequently turned into a political crisis, reflecting deep political divisions on how governments should respond to the pandemic: whether a lockdown is necessary or should governments reopen the economy, and even whether requiring mask-wearing is legitimate. As controversy escalated, Twitter has deleted a post by Trump's personal attorney Rudy Giuliani, endorsing hydroxychloroqine as an effective remedy against coronavirus, and later began affixing fact-check links to tweets by President Trump.[27] In response, the President signed an Executive Order entitled "Preventing Online Censorship." The order claims that online platforms function as "a 21st century equivalent of the public square," accusing them of engaging in "selective censorship" which is harming public discourse, and instructing federal agencies to take action to protect against such alleged censorship.[28] Yet,

---

[21] Keller, *supra* note 5, at 2–3; *see also* Thomas E. Kadri & Kate Klonick, *Facebook v.* Sullivan*: Public Figures and Newsworthiness in Online Speech*, 93 S. CAL. L. REV. 37, 37–40 (2019); Daniel Kreiss & Shannon C. Mcgregor, *The "Arbiters of What Our Voters See": Facebook and Google's Struggle with Policy, Process, and Enforcement Around Political Advertising*, POL. COMM., June 19, 2019, at 2; Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 STAN. TECH. L. REV. 473, 477 (2016); Moran Yemini, *Missing in "State Action": Toward a Pluralist Conception of the First Amendment*, 23 LEWIS & CLARK L. REV. 1149, 1176 (2020).

[22] *See infra* Part III.C.

[23] Kadri & Klonick, *supra* note 21, at 38; Perel & Elkin-Koren, *supra* note 21, at 481.

[24] Michelle Toh, *Facebook, Google and Twitter Crack Down on Fake Coronavirus 'Cures' and Other Misinformation*, CNN BUS. (Feb. 3, 2020, 4:11 AM), https://www.cnn.com/2020/01/31/tech/facebook-twitter-google-coronavirus-misinformation/index.html.

[25] *Id.*

[26] Jack Goldsmith & Andrew Keane Woods, *Internet Speech Will Never Go Back to Normal*, ATLANTIC (Apr. 25, 2020), https://www.theatlantic.com/ideas/archive/2020/04/what-covid-revealed-about-internet/610549/.

[27] Amanda Seitz & Barbara Ortutay, *Tech Companies Step up Fight Against Bad Coronavirus Info*, ASSOCIATED PRESS (Apr. 15, 2020), https://apnews.com/88ccd8d2714998cb06fb88639c271af6.

[28] Exec. Order No. 13925, Preventing Online Censorship, 85 Fed. Reg. 34079 (May 28,

while the Order presumably seeks to create better incentives for platforms to avoid biased restrictions of allegedly legitimate content, it ultimately leaves the decision-making about what speech accounts as legitimate at the hands of the platforms. As a result, it fails to establish a real and meaningful check over the manner in which platforms act as public enforcers of online speech.

Regardless of where one actually draws the line between public actions and private business, the distinction between the two is essential for determining the constitutional analysis. The current body of literature on content moderation by platforms assumes that it is technically feasible to separate the public functions executed by platforms from the private ones. To the extent that platforms exercise governmental powers in allowing or banning speech or speakers, platforms should be subject, as the argument goes, to public law principles of accountability, legitimacy, oversight, and power separation.

In this Paper, we question this approach. We argue that with the increasing use of Artificial Intelligence (AI) by platforms for content moderation, we can no longer distinguish between public functions and private functions executed by platforms. Specifically, in content moderation by AI, the public law enforcement functions are integrated with the private business functions that are driven by commercial interests.[29] The same technical design which is used for targeted advertising and for curating personalized content is also deployed for monitoring and censoring online content.[30] Using machine learning (ML), the system is informed by the same labeling of users and content and makes use of the same application programming interfaces (APIs),[31] learning patterns, and software. Consequently, decisions on removal

2020) (the order instructs several federal agencies to take actions that threaten to limit the legal immunity of platforms for user generated content and jeopardize the economic strength of platforms). Specifically, it directs the Commerce Department to petition the FCC to generate rulemaking implementing a narrower interpretation of Section 230; it directs the Attorney General to prepare alternative legislation; and it instructs federal agencies to review and report their spending in social media advertising. Legal scholars have raised serious doubts as to effective legal power of the executive order, arguing that the FCC, which is an independent federal agency, holds no jurisdiction over rulemaking authority on Section 230. *See* Eric Goldman, *Trump's "Preventing Online Censorship" Executive Order Is Pro-Censorship Political Theater*, TECH. & MKTG. L. BLOG (May 29, 2020), https://blog.ericgoldman.org/archives/2020/05/trumps-preventing-online-censorship-executive-order-is-pro-censorship-political-theater.htm; Jim Wilson, *Explaining President Trump's Executive Order Targeting*, N.Y. TIMES (May 28, 2020) (quoting Ellen Goodman), https://www.nytimes.com/2020/05/28/us/politics/trump-twitter-explained.html. Moreover, as we demonstrate in Part II below, private actors are not bound by the constitutional principles of accountability, transparency, legitimacy, and rational decision-making.

29 *See infra* Part III.C.

30 *See infra* Part III.C.

31 Take, for instance, the way Facebook collects data from the use of Facebook to authenticate identity: "[W]hen users authenticate to websites or applications using their Facebook identities, the API records these acts to their Facebook data profiles. Having access to this identity, many applications then silently contribute to the Facebook social graph via the API, extracting data from our shopping habits or information-seeking behavior and sending it along. Facebook then uses these data traces to tailor advertising and adjust newsfeed priorities, among other customizations to our personalized walled gardens." Jean-Christophe Plantin et al., *Infrastructure Studies Meet Platform Studies in the Age of Google and Facebook*, 20 NEW MEDIA & SOC'Y 293, 304 (2016).

of speech, for (public) law enforcement purposes are driven by the same data, algorithms, and optimization logic, which are also underlying all other functions performed by the platform. As a result, where content moderation is pursued by a single, inextricable system of AI, the public/private classifications largely lose their distinctive power. Therefore, the use of AI in content moderation requires a fresh approach to restraining the power of platforms and securing fundamental freedoms in this environment.

This Paper explores how the legal divide between public and private could be translated into a technological feature. It contends that the normative distinctions between public and private functions could be upheld in online content moderation, provided that these distinctions are embedded in the system design. Rather than simply asking *what* type of power is exercised (public/private), in the era of AI we should also be asking *how* this power is exercised and design our legal and technological remedies accordingly. Understanding the different functions of content moderation by AI and acknowledging their internal independence may offer important insights on what should be done to ensure a check on content moderation by platforms and to subject the public functions of platforms to constitutional restraints.

The Paper proceeds as follows. Part II provides the legal framework for the discussion. It briefly introduces the constitutional framework which is based on the public/private divide, focusing on three of its practical aspects: (1) assuring the nondelegation of governmental powers; (2) subjecting governmental actions to constitutional muster under the state action doctrine; and (3) facilitating checks and balances by ensuring the separation of public powers. Part III describes the rise of content moderation by platforms and demonstrates how it challenges the longstanding constitutional divide between public and private from all three aspects discussed in Part II. Next it maps the different functions performed by platforms in content moderation and analyzes the challenges involved in applying the constitutional framework to these functions. Part IV takes a system perspective and explains how the use of AI for content moderation leads to a fusion of public and private functions. The different functions of content moderation become inextricable because they are all intertwined and embedded in a single technological design.

Finally, Part V proposes a different approach to restraining the power of platforms while enhancing the accountability of content moderation by AI, namely "separation of functions." To separate public functions in an AI-driven private system, it is necessary to take a design-based approach. We propose to separate the platforms' data collection and labeling from the technical tools which are designed to perform the public functions based on that data. The idea is to facilitate independent tools embedding public policy. These tools would run on the platforms' data and would include their own optimization processes informed by public policy. This would enable a dynamic process of adjusting the content moderation algorithm by constant learning. Such separation between independent public tools and private data may enhance public scrutiny of law enforcement speech restrictions, which are traditionally exclusively a public function. This functional separation may also facilitate competition among different players who may enrich the design of speech regulation and mitigate biases. Finally, we explore the implications of this approach and discuss its potential limitations.

II. DEMOCRACY AND THE PUBLIC/PRIVATE DIVIDE

A fundamental premise in the American democratic system is that public and private are distinct spheres.[32] This distinction is considered a necessary pre-condition for liberty itself because it defines a sphere of private activity as sacred, to be free from government intervention.[33] Under constitutional law, it is the exercise of governmental authority that must be accountable to the electorate and subject to the rule of law,[34] because the government has a unique capacity to coerce behavior and undermine individual freedom.[35] By restraining the power of the government and assuring governmental agencies protect civil rights, democracies safeguard against tyranny.[36]

For the purpose of legal scrutiny, most scholars agree that there ought to be a meaningful difference between public and private, and that constitutional restraints should apply only to the former.[37] In fact, "no matter how blurred the line between public and private and no matter how difficult to design an intellectually defensible test to distinguish them,"[38] the public/private divide seems to retain its governing status in constitutional law.

Also, in the area of online content moderation, scholars continue to hold on to the idea that the ways platforms govern online speech should be treated as "public" in their nature, and thus be subject to constitutional-like restraints, such as accountability, transparency, and legitimacy.[39] Nevertheless, the deployment of AI for content moderation is a game changer in this respect, because it integrates public

---

32 Gillian E. Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1369–70 (2003).

33 *See, e.g.*, Lugar v. Edmondson Oil Co., 457 U.S. 922, 936 (1982) (holding that the state action doctrine "preserves an area of individual freedom by limiting the reach of federal law and federal judicial power"). For scholarly articulations of this defense of the state action doctrine, see Charles Fried, *The New First Amendment Jurisprudence: A Threat to Liberty*, 59 U. CHI. L. REV. 225, 229 (1992), Robert H. Mnookin, *The Public/Private Dichotomy: Political Disagreement and Academic Repudiation*, 130 U. PA. L. REV. 1429, 1429 (1982), and Maimon Schwarzschild, *Value Pluralism and the Constitution: In Defense of the State Action Doctrine*, 1988 SUP. CT. REV. 129, 132 (1988).

34 Jody Freeman, *Private Parties, Public Functions and the New Administrative Law*, 52 ADMIN. L. REV. 813, 814 (2000).

35 *See* Yemini, *supra* note 21, at 1170 (discussing the "libertarian premise" that government has such a unique capacity.

36 *See, e.g.*, THE FEDERALIST NO. 51, at 291–92 (James Madison) (Clinton Rossiter ed., 1961) (Madison discusses the way a republican government can serve as a check on the power of factions, and the tyranny of the majority).

37 Freeman, *supra* note 34, at 842.

38 *Id.*; *see also* Metzger, *supra* note 32, at 1369 (citation omitted) (explaining that "private actors are so deeply embedded in governance that 'the boundaries between the public and private sectors' have become 'pervasively blurred'").

39 Kadri & Klonick, *supra* note 21, at 92, 96–97 (arguing that when platforms apply judicial concepts related to newsworthiness and public figures, they act as legislature, executive, judiciary, and press, and therefore, they must separate their powers and create institutions like the Supreme Court to provide transparent decisions and submit to consistent rationales); *see also* Perel & Elkin-Koren, *supra* note 21, at 485–86 (claiming that when platforms perform law enforcement duties, like removing allegedly infringing content upon notice to enjoy the safe harbor under the Digital Millennium Copyright Act, they perform governmental functions and therefore must be held accountable).

functions and private ones in a single, complex technological design that cannot be broken down into distinct and independent functions.[40]

Before this Paper explains why this longstanding divide between public and private can hardly be sustained in online content moderation by AI, it introduces the public/private divide as a basic foundation of the American constitutional framework. This Part focuses on the centrality of this divide to fundamental concepts in constitutional law: limiting the delegation of powers, subjecting governmental conduct to constitutional muster, and facilitating checks and balances through the separation of powers.

A. Limiting the Delegation of Powers

The longstanding divide between public and private is central to the issue of privatization: there are specific public powers that could never be delegated to private actors. The nondelegation doctrine, for instance, sets limits on congressional power to delegate its legislative power directly to the *private* sector.[41] The private nondelegation doctrine forbids the transfer of public power to private entities.[42] Lawmaking has long been considered as "the most important power created for our government by the Founders" since it is "linked to the will of the people through the electoral process and other means."[43] Private actors, to the contrary, are unelected and not sufficiently bound by constitutional principles of accountability, transparency, and legitimacy.[44] As James Boyle explains, lawmaking by private entities raises the dangers of corruption and arbitrariness, and, beyond that, it blurs "the line between public and private, so that public sovereignty would be gifted to private parties, perhaps for populist, redistributive, or simply commercial, rent-seeking ends."[45] Private actors may not adequately represent the public interest but instead be biased towards their own commercial interests.[46]

Private delegations of government power are not bound by constitutional review.[47] Yet, when private actors perform traditionally public functions "unfettered by the scrutiny that normally accompanies the exercise of public power," they may raise accountability concerns "that dwarf the problem of unchecked agency

---

[40] *See infra* Part IV.

[41] Brown, *supra* note 9, at 660 (arguing that the nondelegation doctrine warrants constitutional scrutiny of executive branch outsourcing of legislative power to private parties).

[42] Carter v. Carter Coal Co., 298 U.S. 238, 311 (1936) (holding that private groups—for example, trade or industrial organizations—cannot be empowered to make law).

[43] Scott R. Furlong & Cornelius M. Kerwin, *Interest Group Participation in Rule Making: A Decade of Change*, 15 J. PUB. ADMIN. RES. & THEORY 353, 354 (2005).

[44] The Supreme Court has warned against the risks of delegating powers to a private party "whose interests may be and often are adverse to the interests of others in the same business." *Carter*, 298 U.S. at 311 (striking down statute authorizing local coal boards to determine coal prices and employee wages and hours, based on the Commerce and Due Process Clauses).

[45] Boyle, *Nondelegation Doctrine*, *supra* note 5, at 14.

[46] Harold J. Krent, *Fragmenting the Unitary Executive: Congressional Delegations of Administrative Authority Outside the Federal Government*, 85 NW. U. L. REV. 62, 63–65 (1990) (claiming that congressional delegations outside of the federal government are inconsistent with the separation of powers doctrine as expressed by the Supreme Court).

[47] Metzger, *supra* note 32, at 1370.

discretion."[48] As Jody Freeman argues, the federal judiciary could use the nondelegation doctrine to invalidate private delegations, "especially if the delegated authority implicates 'core' public powers."[49] Indeed, "the powers exercised by private actors as a result of privatization often represent forms of government authority," and, in a sense, "a core dynamic of privatization is the way that it can delegate government power to private hands."[50] Nevertheless, the normative values underlying the structural Constitution—including accountability, transparency, legitimacy, and rational decision-making—do not readily apply to the full spectrum of public-private relationships implicating the exercise of legislative powers.[51]

B. Constitutional Muster

The public/private divide is also important for setting the applicable standard for judicial scrutiny. Generally, the government's use of private sources to conduct its work evades the doctrinal scrutiny that would normally operate to preserve constitutional values.[52] As explained by the Court, the Fourteenth Amendment "affords no shield" against private conduct, "no matter how unfair that conduct may be."[53] In addition, "[t]he primary means available for keeping private actors who exercise public functions within constitutional constraints is the state action doctrine."[54] And "the usual linchpin for finding state action is identifying substantial governmental involvement in the specific private acts being challenged."[55] When such involvement is found, courts may treat private actors as public ones and subject them to the same oversight mechanisms and procedural controls that apply to agents, such as accountability to an elected body and vulnerability to judicial review.[56]

Nevertheless, convincing a court to treat a private actor as a public one for the purpose of constitutional liability is rather challenging.[57] Indeed, private entities and governments "could pursue the most efficient and effective forms of program" when they are "unconcerned with constitutional requirements."[58] Many times, private parties' involvement in governance is "an indirect side-effect of their autonomous determinations, often made pursuant to independent professional

---

48 Freeman, *supra* note 34, at 818.

49 Jody Freeman, *The Private Role in Public Governance*, 75 N.Y.U. L. REV. 543, 584 (2000).

50 Metzger, *supra* note 32, at 1396.

51 Kimberly N. Brown, *"We the People," Constitutional Accountability, and Outsourcing Government*, 88 IND. L.J. 1347, 1369, 1376 (2013).

52 *See* Jack M. Balkin, *Respect-Worthy: Frank Michelman and the Legitimate Constitution*, 39 TULSA L. REV. 485, 486 (2004).

53 Nat'l Collegiate Athletic Ass'n v. Tarkanian, 488 U.S. 179, 191 (1988).

54 Brown, *supra* note 9, at 627; *see also* BeVier & Harrison, *supra* note 12 (discussing the merits and criticisms of the state action doctrine).

55 Metzger, *supra* note 32, at 1370.

56 Freeman, *supra* note 34, at 819, 842.

57 *See* Evans v. Newton, 382 U.S. 296, 299–300 (1966); Metzger, *supra* note 32, at 1421; *see also* Louis Michael Seidman, *The State Action Paradox*, 10 CONST. COMMENT. 379, 391 (1993) ("No area of constitutional law is more confusing and contradictory than state action.").

58 Metzger, *supra* note 32, at 1401.

standards."[59] Hence, the broad discretion often granted to private actors can support seeing their underlying conduct "as being one that the government seeks to foster independent private action."[60] Applying constitutional norms to independent private action could intrude on private autonomy and private actors' freedom to act as they see fit.[61] Hence, even if private actors exercise power over vulnerable third parties and control access to public goods, this might not be enough to distinguish the powers exercised as governmental.[62]

C. Checks and Balances

Separation of powers is frequently portrayed as the unique genius of the United States Constitution, the very basis for the success of American democracy.[63] The constitutional theory of checks and balances provides practical security against the excessive concentration of political power in one branch of the government.[64] It assumes that unlimited power is likely to be misused and encroach on individual liberties.[65] By giving "those who administer each department the necessary constitutional means and personal motives to resist encroachments of the others," the Framers sought to create a system in which competition among the branches would limit overreach by any one of them—in which "[a]mbition [would] be made to counteract ambition."[66] The idea is that "[i]f one branch fell under the control of a would-be monarch or tyrannical cabal, the other branches might provide a check by using their constitutional powers to block oppressive measures."[67]

In the United States, governmental power is divided between three branches, created and supported by constitutional values.[68] Congress and state legislatures make laws, the executive branches enforce those laws, and courts validate their legality against the Constitution.[69] The role of media as a watchdog of the government was often informally referred to as the "fourth branch."[70] Each branch was made "answerable to different sets of constituencies and subject to different temporal demands."[71] Institutionalizing such a differentiation between executive, legislative,

---

[59] *Id.* at 1398.

[60] *Id.* at 1399–1400.

[61] *Id.* at 1406.

[62] *Id.* at 1398.

[63] *See, e.g.*, Hugh Heclo, *What Has Happened to the Separation of Powers?*, *in* SEPARATION OF POWERS AND GOOD GOVERNMENT 131, 133–34 (Bradford P. Wilson & Peter W. Schramm eds., 1994).

[64] Daryl J. Levinson & Richard H. Pildes, *Separation of Parties, Not Powers*, 119 HARV. L. REV. 2311, 2316–17 (2006).

[65] *See* THE FEDERALIST NO. 51, at 291–92 (James Madison) (Clinton Rossiter ed., 1961).

[66] Jeffrey A. Love & Arpit K. Garg, *Presidential Inaction and the Separation of Powers*, 112 MICH. L. REV. 1195, 1203–04 (2014) (citation omitted).

[67] Levinson & Pildes, *supra* note 64, at 2319.

[68] Kadri & Klonick, *supra* note 21, at 93.

[69] *See* Marbury v. Madison, 5 U.S. (1 Cranch) 137 (1803).

[70] Rachel Luberda, *The Fourth Branch of the Government: Evaluating the Media's Role in Overseeing the Independent Judiciary*, 22 NOTRE DAME J.L. ETHICS & PUB. POL'Y 507, 508 (2008).

[71] Jon D. Michaels, *An Enduring, Evolving Separation of Powers*, 115 COLUM. L. REV. 515, 525 (2015).

and judicial powers is expected to "harness political competition into a system of government that would effectively organize, check, balance, and diffuse power."[72] This system was envisioned as "a machine that would go of itself,"[73] relying on "interbranch competition to police institutional boundaries and prevent tyrannical collusion."[74]

This public law principle of separation of power, however, seems to be evolving.[75] It is not only the government that triggers the commitment to check, separate, and balance its powers, but any exercise of state power.[76] Administrative power, for instance, is divided "among politically appointed agency leaders, an independent civil service, and a vibrant civil society."[77] This administrative separation of power,[78] some argue, should even extend beyond the administrative state.[79] How far beyond? This, of course, depends on where we draw the line between governmental functions and private ones.

All in all, while the line between public and private functions is becoming more complicated to draw, it remains the touchstone of constitutional law. Private delegations—no matter how bothering the way they impact civil liberties could be—escape "a handful of baseline values for good government" which "influence the exercise of public power at the governmental end of the continuum."[80] It is only governmental actions that are subject to constitutional scrutiny and restrained by the separation of powers to assure their legitimacy and safeguard from abuse of power. As we show next, sustaining this longstanding divide in the context of content moderation by platforms is extremely difficult. When platforms deploy AI for content moderation, it even becomes futile.

III. CONTENT MODERATION BY PLATFORMS: A CONSTITUTIONAL CHALLENGE

The constitutional public-private divide is challenged by the rise of platforms as a major force that dominates our public sphere. The public sphere, where people can gain access to information, exchange ideas and knowledge, establish their opinions, and develop their identities, is a fundamental tenet of democracy.[81] The ideal

---

72 Levinson & Pildes, *supra* note 64, at 2313.

73 MICHAEL KAMMEN, A MACHINE THAT WOULD GO OF ITSELF: THE CONSTITUTION IN AMERICAN CULTURE 17–18 (1987).

74 Levinson & Pildes, *supra* note 64, at 2313.

75 Michaels, *supra* note 71, at 520–21.

76 *Id.* at 517.

77 *Id.* at 520; Elizabeth Magill & Adrian Vermeule, *Allocating Power Within Agencies*, 120 YALE L.J. 1032, 1035 (2011).

78 *See* Neal Kumar Katyal, *Internal Separation of Powers: Checking Today's Most Dangerous Branch from Within*, 115 YALE L.J. 2314, 2343 (2006); Gillian E. Metzger, *The Interdependent Relationship Between Internal and External Separation of Powers*, 59 EMORY L.J. 423, 428–29 (2009); Michaels, *supra* note 71, at 520.

79 Michaels, *supra* note 71, at 535.

80 Brown, *supra* note 9, at 618.

81 Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017); *see also* Niva Elkin-Koren, *Cyberlaw and Social Change: A Democratic Approach to Copyright in Cyberspace*, 14 CARDOZO ARTS &

of self-governance by the people in liberal democracies assumes the free flow of information and deliberation of the governing people in the public sphere.[82]

Nowadays, our public sphere is dominated by online platforms.[83] The distributed design of the internet, which connected users and content via distributed networks, is now mediated by mega platforms, such as Facebook, Twitter, and Google. Due to strong network effects, these platforms effectively govern online access to content and speakers and control the proliferation of online expressions. Intersecting the voluminous flows of online content and matching between expressions and potential audience, platforms offer a natural point of control for monitoring, filtering, blocking, and disabling access to online content. Platforms may enable or disable access to content by removing or blocking controversial content or by terminating the accounts of particular speakers.[84] This gatekeeping function has also made platforms ideal partners for performing civil and criminal law enforcement.[85]

Platforms are far more than a neutral infrastructure that connects users[86] and enables the sharing of User-Generated Content (UGC).[87] They are shaping our public discourse in varied ways.[88] Platforms define what content can be uploaded and shared (e.g., Facebook Community Standards),[89] which content would remain available and which would be removed (e.g., hate speech, terrorist propaganda, copyright infringement),[90] who can participate in online conversation (e.g., verifying online identity, suspending accounts),[91] how content might be shared (e.g, "like" or

---

ENT. L.J. 215, 218 (1996).

82 Whitney v. California, 274 U.S. 357, 375–77 (1927) (Brandeis, J., concurring).

83 Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2012–15 (2018); James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L & TECH. 42, 45, 48 (2015); Danielle Keats Citron & Neil M. Richards, *Four Principles for Digital Expression (You Won't Believe #3!)*, 95 WASH. U. L. REV. 1353, 1355–60 (2018); Keller, *supra* note 5, at 1.

84 Keller, *supra* note 5, at 1.

85 Bridy, *supra* note 20, at 83–84; Zittrain, *supra* note 20, at 255–57.

86 Platforms tend to present themselves as simply offering a point of connection, like Facebook connecting people together and Uber connecting car owners and potential customers. *See* Lior Zalmanson & Thomas Gegenhuber, *When Algorithms are Your Boss: Staying Human in Platform-Mediated Work*, RE:PUBLICA: SCI. & TECH. (Feb. 5, 2018), https://18.re-publica.com/de/session/when-algorithms-your-boss-staying-human-platform-mediated-work.

87 Tarleton Gillespie, *Regulation of and by Platforms*, *in* THE SAGE HANDBOOK OF SOCIAL MEDIA 254, 256–58 (Jean Burgess et al. eds., 2018).

88 *See, e.g.*, Keller, *supra* note 5, at 1; Perel & Elkin-Koren, *supra* note 21, at 488.

89 Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (Nov. 15, 2018),https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

90 *See generally* Niva Elkin-Koren & Maayan Perel, *Guarding the Guardians: Content Moderation by Online Intermediaries and the Rule of Law*, *in* OXFORD HANDBOOK OF ONLINE INTERMEDIARY LIABILITY 669 (Giancarlo Frosio ed., 2020) (discussing the increasing pressure put on online platforms to block, remove, and monitor illegitimate content).

91 Paul Hildago, *Call for 'Universal Verification' on Social Media*, THE HILL (Nov. 28, 2018, 2:00 PM), https://thehill.com/opinion/technology/418159-call-for-universal-verification-on-social-media.

"retweet")[92] and who is likely to watch it (e.g., YouTube recommendation system).[93]

By the nature of their business, platforms essentially stand between potential speakers and their potential audience in ways that traditionally only governments could and, in fact, in many ways which governments never could.[94] Still, however, the constitutional divide between public and private has been so far unsuccessful in bounding speech regulation by online platforms to constitutional restraint.[95] Despite concerns that "the real threat to free speech today comes from private entities such as Internet service providers, not from the Government,"[96] interfering with the editorial discretion of platforms is seen as a violation of platforms' own First Amendment rights.[97] Essentially, requiring platforms to host content against their will arguably forces them to speak, in violation of the First Amendment.[98] Thus, although the social web is perhaps the place "where the line between public and private seems least clear,"[99] when it comes to constitutional values, it is firmly treated as a private sphere.

In the following discussion, we use the fundamental concepts described in Part II to show that contrary to any dichotomous vision of the public/private divide, content moderation by platforms simply does not fall neatly within these constitutional categories.

A. Privatization

Content moderation by online platforms has been repeatedly addressed as a case of privatization.[100] Many scholars have shown "how governments can bypass

---

92 Plantin et al., *supra* note 31, at 7.

93 Már Másson Maack, *'YouTube Recommendations Are Toxic,' Says Dev Who Worked on the Algorithm*, THENEXTWEB (June 14, 2019, 3:58 PM), https://thenextweb.com/google/2019/06/14/youtube-recommendations-toxic-algorithm-google-ai/.

94 Moran Yemini, *The New Irony of Free Speech*, 20 COLUM. SCI. & TECH. L. REV. 119, 119–22 (2018).

95 *See, e.g.*, Order Granting Defendant's Motion to Dismiss at 16–17, Prager Univ. v. Google (2018) (No. 17-CV-06064-LHK), 2018 WL 1471939 (rejecting conservative commentator Dennis Prager's claim that YouTube violated the First Amendment when it limited users' access to his videos).

96 U.S. Telecom Ass'n v. FCC, 855 F.3d 381, 434 (D.C. Cir. 2017) (per curiam) (Kavanaugh, J., dissenting).

97 Keller, *supra* note 5, at 2.

98 La'Tiejira v. Facebook, Inc., 272 F. Supp. 3d 981, 991–92 (S.D. Tex. 2017); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 630 (D. Del. 2007); Search King, Inc. v. Google Tech., Inc., No. CIV-02-1457-M, 2003 WL 21464568 (W.D. Okla. Aug. 7, 2003); Daniel (Yue) Zhang et al., *Crowdsourcing-Based Copyright Infringement Detection in Live Video Streams*, *in* PROC. OF THE 2018 IEEE/ACM INTERNATIONAL CONFERENCE ON ADVANCES IN SOCIAL NETWORKS ANALYSIS & MINING (2018).

99 Sarah Michele Ford, *Reconceptualizing the Public/Private Distinction in the Age of Information Technology*, 14 INFO. COMM. & SOC'Y 550, 558 (2011).

100 Boyle, *Nondelegation Doctrine*, *supra* note 5, at 10; Perel & Elkin-Koren, *supra* note 21, at 485; Sebastian Felix Schwemer, *Trusted Notifiers and the Privatization of Online Enforcement*, 35 COMPUTER L. & SEC. REV., no. 6, 2019, at 19. *See generally* Wendy Seltzer, *Free Speech Unmoored in Copyright's Safe Harbor: Chilling Effects of The DMCA on The First Amend.*, 24 HARV. J.L. & TECH. 171 (2010).

constitutional limits by deputizing private platforms as censors."[101] For instance, to address copyright enforcement in the digital age, the government, by enacting the U.S. Digital Millennium Copyright Act (DMCA), ultimately pushed platforms to create a private police force against online infringement that is not bound by statutory and constitutional privacy constraints.[102] Obviously, such intermediation, which is inherent to secondary liability regimes,[103] is not the "paradigm case" tackled by the non-delegation doctrine—the inexplicable delegation of legislative tasks by Congress to an administrative agency.[104] Rather, this is a form of private delegation. Nevertheless, whether deployed by private bodies empowered by government or by "'neutral technology' backed by government standard-setting powers," it seems like an exercise of public power that "should not escape completely from the world of democratic and constitutional review."[105]

Indeed, platforms effectively exercise governmental powers when they elaborate rules and systems to resolve collisions between preserving free expression and regulating harmful speech.[106] It has been argued that these rules are made and enforced in ways that are comparable to actual legislation, and they evolve in ways that are similar to common-law judicial adjudication.[107] For instance, Facebook's determination that the phrase "[s]omeone shoot Trump" should be deleted because the U.S. President is a "protected category," but the sentence "[t]o snap a bitch's neck, make sure to apply all your pressure to the middle of her throat" should not be seen as a credible threat,[108] is a *de facto* exercise of rulemaking power.

Sometimes content is removed or blocked when it is contrary to the platforms' terms of use. Indeed, platforms may opt to make content moderation decisions based on their terms of service, rather than the law of the land, to strengthen their legal discretion over removal decisions.[109] In other cases platforms would opt to

---

101 Keller, *supra* note 5, at 2; Elkin-Koren & Haber, *supra* note 6, at 107; Meyerson, *supra* note 8, at 118 (referring to the phenomenon as "collateral censorship"); Kreimer, *supra* note 7, at 16 (calling it "censorship by proxy").

102 *See* James Boyle, *A Politics of Intellectual Property: Environmentalism for the Net?*, 47 DUKE L.J. 87, 100–07 (1997); James Boyle, *Intellectual Property Policy Online: A Young Person's Guide*, 10 HARV. J.L. & TECH. 47, 101–05 (1996).

103 Seltzer, *supra* note 100, at 181.

104 Boyle, *Nondelegation Doctrine*, *supra* note 5, at 13.

105 *Id.* at 16.

106 *See generally* Klonick, *supra* note 3, at 1630–58 (discussing how platforms are "governing" through "private content moderation systems").

107 *See* Kate Klonick, *Facebook v. Sullivan*, KNIGHT FIRST AMEND. INST., Oct. 1, 2018, at 6. *But see* David Pozen, A*uthoritarian Consitutiutionalism in Facebookland* (October 20, 2018) available at https://knightcolumbia.org/content/authoritarian-constitutionalism-facebookland (arguing that unlike common law system, Facebook's content moderation regime lacks formally independent dispute resolution bodies, and since regulators and adjudicators are one and the same it should be viewed more like a system of authoritarian constitutionalism).

108 *Facebook's Manual on Credible Threats of Violence*, GUARDIAN (May 21, 2017, 1:00 PM), https://www.theguardian.com/news/gallery/2017/may/21/facebooks-manual-on-credible-threats-of-violence; *see also* Jon Fingas, *Facebook Defends Content Policy After Guidelines Leak*, ENDGADGET (May 23, 2017), https://www.engadget.com/2017/05/23/facebook-defends-content-guidelines.

109 *See, e.g.*, BEN WAGNER, GLOBAL FREE EXPRESSION–GOVERNING THE BOUNDARIES OF

remove content that is illegal (for instance, hate crime and violent threats). In such cases platforms are applying their own interpretation of criminal laws, namely where to draw the line between legitimate exercise of freedom of expression and speech that might constitute criminal conduct.[110] The interpretation of criminal law must be carried out in light of affected constitutional rights, a role that is reserved to courts. Delegating the responsibility to remove illegal content to private entities also delegates the application and interpretation of legal norms, which is essentially the role of government.[111]

Nonetheless, in the United States, content moderation as a form of private delegation is not bound by constitutional restraint.[112] After all, online platforms remain non-federal actors.[113] Notwithstanding how powerful they have become in mediating our public sphere,[114] they are profit-maximizing businesses. In any event, even if "the private nondelegation doctrine could" presumably "play an important role in encouraging greater scrutiny over" content moderation by platforms,[115] as we explain in Part IV, the use of AI for content moderation seriously inhibits this possibility. Specifically, when lawmaking power is inextricably tied to the legitimate exercise of private business-related discretion,[116] it becomes extremely complicated to bind content moderation as a whole by traditional constitutional principles of accountability and legitimacy.

B. Constitutional Scrutiny

It is undisputed that in the digital ecosystem, the role of states and corporations and the consequences of their actions have converged.[117] Nevertheless, it seems like the public/private divide still serves the courts in their efforts to draw the line between speech regulation that is subject to First Amendment scrutiny and discretional content management that is not. For instance, in *Manhattan Community Access Corp. v. Halleck*,[118] the Supreme Court recently held that a TV station run by a private nonprofit corporation is not a state actor and therefore was subject to no duty under the First Amendment.[119] The defendant, Manhattan Neighborhood Network (MNN), a public access television network that serves New York City, refused to

---

INTERNET CONTENT 111–12 (2016).

110 Katharina Kaesling, *Privatising Law Enforcement in Social Networks: A Comparative Model Analysis*, 3 ERASMUS L. REV. 151, 159 (2018).

111 *Id.*

112 Brown, *supra* note 9, at 618; Freeman, *supra* note 34, at 839.

113 *See* Klonick, *supra* note 3, at 1658–62.

114 Balkin, *supra* note 83, at 2011, 2020, 2041; *see* Yemini, *supra* note 94, at 119–20.

115 Boyle, *Nondelegation Doctrine*, *supra* note 5, at 16.

116 *See infra* Part IV.

117 Ford, *supra* note 99; Yemini, *supra* note 21, at 1172.

118 Manhattan Cmty. Access Corp. v. Halleck, No. 17-1702, slip op. (587 U.S. __ June 17, 2019).

119 *Id.* at 2. Indeed, this does not mean that in this capacity social media platforms are immune from any governmental regulation. It is a separate question whether social media platforms are protected against governmental regulation which may require them to enable the speech of others. Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 636–37 (1994).

broadcast a video by DeeDee Halleck, an award-winning producer, and Jesus Papoleto Melendez, a poet and playwright, claiming it contained threatening language. The plaintiffs argued that MNN's actions violated their First Amendment rights,[120] but the Supreme Court concluded that MNN was not a state actor subject to the First Amendment, explaining that a private entity may qualify as a state actor when it exercises "powers traditionally exclusively reserved to the State,"[121] and "[p]roviding some kind of forum for speech is not an activity that only governmental entities have traditionally performed."[122] Similarly, it was held that the mere fact that a private entity is subject to regulatory control, does not in itself justify subjecting the private entity to the constraints of the First Amendment.[123]

Courts applied the same line of reasoning to declare, over and over again, that online platforms are not bound by the First Amendment.[124]

But, merely focusing on the characterization of activity as essentially public or private seems formalistic and conceptually dubious.[125] As we explained, platforms are not just neutral infrastructures for connecting users and sharing content.[126] They arguably perform governmental functions too, which affects the freedom of expression of their users and subsequently affects free speech of the public at large.[127] Platforms effectively assume a state-like role in managing individuals' rights, thereby effectively acting as "private" regulators of public space.[128] Scholars have therefore

---

120 *Halleck*, slip op. at 3–4.

121 *Id.* at 6 (quoting Jackson v. Metro. Edison Co., 419 U.S. 345, 352 (1974)).

122 *Id.* at 9.

123 *Id.* at 13 (citing Denver Area Educ. Telecomm. Consortium, Inc. v. FCC, 518 U.S. 727, 829 (1996)) ("Not surprisingly, as Justice Thomas has pointed out, this Court has 'never even hinted that regulatory control, and particularly direct regulatory control over a private entity's First Amendment speech rights,' could justify subjecting the regulated private entity to the constraints of the First Amendment.").

124 *See, e.g.*, Howard v. Am. Online Inc., 208 F.3d 741, 754 (9th Cir. 2000) (holding that AOL is not a "quasi-public utility" and not a state actor); Murawski v. Pataki, 514 F. Supp. 2d 577, 588 (S.D.N.Y. 2007) (holding Yahoo! could not be held accountable for censoring political messages); Noah v. AOL Time Warner, Inc., 261 F. Supp. 2d 532, 546 (E.D. Va. 2003) (holding that account termination, even if done simply to suppress speech, does not violate the First Amendment because AOL is not a state actor); Island Online, Inc. v. Network Sol., Inc., 119 F. Supp. 2d 289, 307 (E.D.N.Y. 2000) (holding that defendant's policy of filtering out certain domain names does not violate the First Amendment); Thomas v. Network Sol., Inc., 176 F.3d 500, 508 (D.C. Cir. 1999) (holding domain name assignment is not state action); Sanger v. Reno, 966 F. Supp. 151, 163 (E.D.N.Y. 1997) (holding that "Internet providers are not state actors" and are, therefore, "free to impose content-based restrictions on access to the Internet without implicating the First Amendment"); Cyber Promotions, Inc. v. Am. Online, Inc., 948 F. Supp. 436, 437, 452 (E.D. Pa. 1996) (refusing to conduct a First Amendment analysis of AOL's policy against "junk" e-mail because AOL is not a state actor); *cf.* Young v. Facebook, Inc., 790 F. Supp. 2d 1110, 1116 (N.D. Cal. 2011) (holding that account termination by Facebook is not reachable by the First Amendment).

125 Freeman, *supra* note 34, at 842.

126 *See infra* Part III.B.

127 *See* Woodhull Freedom Found. v. United States, 334 F. Supp. 3d 185 (D.D.C. 2018), *rev'd*, 948 F.3d 363 (D.C. Cir. 2020).

128 Keats Citron & Richards, *supra* note 83, at 1361–64; Gregory P. Magarian, *Forward into the Past: Speech Intermediaries in the Television and Internet Ages*, 71 OKLA. L. REV. 237, 238 (2018)

contended that speech regulation by platforms should be treated as state action.[129] Others (including the authors) have previously argued that when platforms perform public functions which were meant to serve the public at large under formal or informal delegation of power from the government, they effectively function like private administrative agencies that should be held accountable for their actions.[130] Nevertheless, even if courts reverse their rejection of free speech-related claims against platforms on the basis of the state action doctrine, two major challenges will remain: determining *which* of these functions are effectively public and determining *how* they could be technically separated in a regime of speech regulation governed by AI. These issues will be discussed in Part IV.

C. Digital Checks and Balances

The convergence of digital powers at the hands of a handful of mega platforms poses an unexplored challenge to the fundamental principle of checks and balances.[131] Specifically, the separation of powers is remarkably absent from content moderation by online platforms.[132] Indeed, "much of the governance of online speech is done by private platforms" that operate in all branches—"legislative, executive, judiciary, and press—at once."[133] Consider YouTube's Content ID as an example.[134] This system was designed to flag content which failed to comply with YouTube's copyright policies. The system enables YouTube to automatically screen user-uploaded content and identify copyrighted content using a digital identifying code.[135] It is also algorithmically set to determine which specific level of similarity between an uploaded video and an original copyrighted work would trigger the matching feature, which will then submit a signal to the right holder, allowing her to choose whether to remove, monetize, block, or disable the allegedly infringing material before it becomes publicly available.[136] YouTube effectively exercises judicial power when it determines which content constitutes an infringement of an original copyrighted work. It also exercises executive power when it acts to remove, disable, or filter such content.[137] As identified by Lisa Bressman, this sort of private lawmaking "interferes with individual liberty for suspect public purposes and

(arguing that "the new intermediaries of the Internet Age operate substantially free of effective regulatory or normative controls").

129 *See* Yemini, *supra* note 21, at 1169. *See generally, e.g.*, DAWN C. NUNZIATO, VIRTUAL FREEDOM: NET NEUTRALITY AND FREE SPEECH IN THE INTERNET AGE (2009).

130 Edward Lee, *Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten*, 49 U.C. DAVIS L. REV. 1017, 1049 (2016); Perel & Elkin-Koren, *supra* note 21, at 483.

131 *See* Niva Elkin-Koren & Maayan Perel, *Algorithmic Governance by Online Intermediaries*, *in* THE OXFORD HANDBOOK OF INTERNATIONAL ECONOMIC GOVERNANCE AND MARKET REGULATION 3 (Eric Brousseau, Jean-Michel Glachant, Jérôme Sgard eds., 2019).

132 Perel & Elkin-Koren, *supra* note 21, at 481.

133 Kadri & Klonick, *supra* note 21, at 93.

134 Perel & Elkin-Koren, *supra* note 21, at 477–78.

135 *Id.*

136 *Id.* at 510.

137 *Id.* at 483.

inadequately reflects a broad public purpose to justify such interference."[138] In other words, the concentration of such public powers at the hands of platforms weakens existing safeguards to freedom of expression embedded in the separation and often competition among the different branches.

Scholars have attempted to use the constitutional public/private divide to introduce checks and balances in content moderation by platforms. For instance, Thomas Kadri and Kate Klonick proposed enabling people to appeal platforms' decisions about their content by establishing a sort of independent "supreme court" that will review these decisions.[139] Similarly, Kyle Langvardt explored the possibility of forming an administrative monitoring and compliance regime to ensure that content moderation policies are in line with First Amendment principles.[140] These scholars emphasized the significance of platforms as public forums for the exchange of views, which have displaced parks and streets.[141] Nevertheless, as we explain henceforth in more detail, platforms are not only public forums. They are—first and foremost—private entities. Thus, even if we could say they are governmental actors to some extent, they are not entirely public in their conduct. Drawing the line between public and private for the purpose of introducing checks and balances to the exercise of governmental power is rather puzzling; as we show next, in Part IV, when AI is involved, it might not even be worth the try, for public functions executed by platforms are deeply integrated with their private functions. Hence, to facilitate the normative values that underline the constitution, including transparency, accountability, and legitimacy, a new approach for maintaining the public/private divide should be considered.

D. The Multiple Functions of Content Moderation

Platforms' use of AI in content moderation could be simultaneously viewed as a practical need to operate in a dynamic, ever-growing digital landscape; as an innovative competitive advantage; and as an expression of responsibility to public values. The different functions of content moderation performed by platforms are situated differently within the traditional public/private divide. To fully understand how AI may blur the distinction between public and private actions, it is important to distinguish between the different types of content moderation performed by platforms using AI.

1. Content Matching Services

Social media platforms use AI to match users and content.[142] Indeed, the public sphere in social media platforms does not exist in the same sense that we conceive it in mass media. It is fragmented into segmented views, where each user receives a curated, personalized view of the entire public discourse that is not

---

138 Lisa Schultz Bressman, Schechter Poultry *at the Millennium: A Delegation Doctrine for the Administrative State*, 109 YALE L.J. 1399, 1428 (2000).

139 Kadri & Klonick, *supra* note 21, at 69, 94.

140 Langvardt, *supra* note 4, at 1353, 1377.

141 *Id.* at 1356.

142 Force v. Facebook, Inc., 934 F.3d 53, 58 (2d Cir. 2019).

necessarily shared by others.[143]

The business model of multisided-platforms[144] is based on generating data on users and extracting revenues from selling users' profiles for targeted advertising or other data driven products and services.[145] Media scholars have shown how the commercial logic of social media platforms is driving their technical design.[146] Advertising revenues and overall revenues from data collection depend on the three V's of data: volume, velocity, and variety. To enhance the amount of data collected on each user, the types of data collected, and the freshness of data, platforms seek to enhance the amount of time and attention spent on the platform.[147] Platforms, therefore, seek to attract users by matching them with content that best fits their preferences.[148]

The dominant power of platforms to decide which content becomes available to which audience is what drives the platforms' earning capacity: the better the match between content and users, the more attractive the services of the platform become. The challenge of every social media platform is to generate as accurate matches as possible so that users will be satisfied with the content they encounter, while, in turn, "surrendering" their valuable trail of personal data for the platforms' economic benefit. If the content that users encounter does not match their personal interests, or is otherwise polluted with disinformation, child pornography, extremist content, or hoaxes, platforms may lose their legitimacy, popularity, and, consequently, lose their earnings.[149]

It is at this point where the unprecedented possibilities of AI come into play. The advanced ability to collect users' data and then apply ML technologies to predict and even shape their personal preferences enables platforms to optimize their matchmaking capabilities.[150]

For instance, relying on deep learning AI, "recommendation systems provide

---

143 Laura Reed & Danah Boyd, *Who Controls the Public Sphere in an Era of Algorithms?*, DATA & SOC'Y 5–6 (May 13, 2016), https://datasociety.net/pubs/ap/QuestionsAssumptions_background-primer_2016.pdf.

144 *See generally* DAVID S. EVANS & RICHARD SCHMALENSEE, MATCHMAKERS: THE NEW ECONOMICS OF MULTISIDED PLATFORMS (2016).

145 Adam C. Uzialko, *How Businesses Are Collecting Data (And What They're Doing with It)*, BUS. NEWS DAILY (Aug. 3, 2018), https://www.businessnewsdaily.com/10625-businesses-collecting-data.html.

146 Plantin et al., *supra* note 31, at 7. Corporations' goal of gathering users' personal data determines the technical properties of platforms, which in turn shapes how they organize communication among users. These affordances are driven by economic interests. "For example, 'like,' 'share,' and 'retweet' not only provide a means for users to express themselves but also facilitate ranking, product recommendations, and data analytics." *Id.*

147 Jack M. Balkin, *Fixing Social Media's Grand Bargain*, 1, 2 (Hoover Institution, Aegis Series Paper No. 1814, 2018), https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf.

148 Old media was also a two-sided market (e.g., newspapers, TV shows) which sought to attract both readers and advertisers.

149 Alexis C. Madrigal, *'The Basic Grossness of Humans'*, ATLANTIC: TECH. (Dec. 15, 2017), https://www.theatlantic.com/technology/archive/2017/12/the-basic-grossness-of-humans/548330/.

150 Force v. Facebook, Inc., 934 F.3d 53, 58–59 (2d Cir. 2019).

a way to suggest similar products (such as in Amazon), news articles (Huffington Post), TV shows (Netflix) or videos (YouTube) to users."[151] To ensure that content is optimized for the intended audience, recommendation systems have "one neural network for gathering user information (such as watch history and user feedback) and another neural network for ranking the selected videos that are displayed."[152] This allows platforms to identify patterns of content preferences that are less obvious.[153] By tailoring specific content to users, platforms keep users logged into the platform, induce the engagement of users with the content, and maximize the time users spend on the platform to optimize the collection of data and exposure to advertising content.[154]

This commercial goal is ultimately shaping which data is collected and how content is organized. YouTube, for instance, would suggest to viewers content that they are likely to continue watching, using AI to predict what they are likely to watch based on the views they have already made.[155] Similarly, Facebook is using data on users' behavior on the platform and elsewhere to adjust the newsfeed priorities of each user.[156]

2. Adjudicating Content

Social media platforms are applying AI-based content moderation to adjudicate conflicting claims regarding the legitimate use of content on their systems. Indeed, illicit, hateful, illegal, or otherwise unwanted or objectionable content might lead to brand degradation.[157] Effectively deploying AI to implement community guidelines and content moderation policies may reduce this risk.[158]

AI content moderation systems are designed to optimize a speedy detection of content that might be considered harmful by claimants.[159] For instance, to benefit from the safe harbor protection offered by the DMCA and expeditiously remove allegedly infringing copyright materials upon receiving a notice from rights holders,[160] platforms have automated their systems of removal upon notice.[161] Today, Content ID can detect and notify rights holders whenever a newly uploaded video matches a work that they own, shifting the detection burden from rights holders to

---

151 Raghav Bharadwaj, *AI for Social Media Censorship–How it Works at Facebook, Youtube, and Twitter*, EMERJ (Feb. 10, 2019), https://emerj.com/ai-sector-overviews/ai-social-media-censorship-works-facebook-youtube-twitter/.

152 *Id.*

153 *Id.*

154 Maack, *supra* note 93.

155 Kevin Roose, *The Making of a YouTube Radical*, N.Y. TIMES (June 8, 2019), https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html.

156 Plantin et al., *supra* note 31, at 16.

157 This was recently demonstrated in the case of the popular online conference application Zoom, whose vulnerable security system was abused to spread obscene materials, the spreading of which could be harmful to the brand. Anu Thomas, *Zoom Turns to AI to Block Nudity on its Platforms*, ANALYTICS INDIA MAG (Apr. 2020) https://analyticsindiamag.com/zoom-turns-to-artificial-intelligence-to-block-nudity-on-its-platform/ (last visited June 11, 2020).

158 *Id.*

159 *See infra* Part IV.A.

160 Digital Millennium Copyright Act of 1998, 17 U.S.C. § 512 (c)–(d) (2012).

161 *See supra* notes 131–35 and accompanying text.

AI.[162] The system incorporates real time data on uploaded content and users. Another example is Scribd, a subscription based digital library of books, which has developed BookID. This system generates a digital fingerprint for each book, based on semantic data (e.g., word count, letter frequency, phrase comparison). Texts uploaded to Scribd are scanned by BookID, and content that matches any BookID fingerprint is blocked.[163]

The use of AI for content adjudication enables platforms to proactively report potentially problematic content to their team of reviewers and even take action on the content automatically.[164] In fact, during the Covid-19 pandemic, major social media platforms, including Facebook,[165] YouTube,[166] and Twitter[167] announced they would shift their content moderation to AI, since their human reviewers were absent due to mandatory lockdowns.[168]

AI not only helps platforms identify and remove a much larger percent of potentially harmful content, but also enables them to remove it faster, before anyone even sees it.[169] According to Google Transparency Report, more than two-thirds of the videos YouTube removed between January and March 2019 were identified automatically, before having any views at all.[170]

Amazon is also using AI to proactively adjudicate apparently illegitimate uses of brands. Powered by Amazon's ML, Project Zero continuously scans Amazon's online stores against key data points that brands provide (e.g., trademarks, logos, etc.), and proactively removes suspected counterfeits before they reach a customer.[171] Facebook too is focusing on improving its ability to detect hate speech, developing a self-supervised AI-building approach, which promises to "help the social network spot the offensive content in its ever-changing forms."[172] Unlike Facebook's older AI systems, which rely on a supervised learning approach of taking

---

162 John Paul Titlow, *YouTube is Using AI to Police Copyright—to the Tune of $2 Billion in Payouts*, FAST COMPANY (July 13, 2019), https://www.fastcompany.com/4013603/youtube-is-using-ai-to-police-copyright-to-the-tune-of-2-billion-in-payouts.

163 *Copyright, DMCA, and BookID*, SCRIBD, https://www.scribd.com/copyright/bookid (last visited Mar. 28, 2020).

164 Zuckerberg, *supra* note 89.

165 *Keeping Our People and Our Platforms Safe*, FACEBOOK NEWS (Mar. 16, 2020, 8:46 PM), https://about.fb.com/news/2020/04/coronavirus/#keeping-our-teams-safe.

166 *Protecting our Extended Workforce and the Community*, YOUTUBE: CREATOR BLOG (Mar. 16, 2020), https://youtube-creators.googleblog.com/2020/03/protecting-our-extended-workforce-and.html.

167 Vijaya Gadde & Matt Derella, *An Update on our Continuity Strategy During COVID-19*, TWITTER: BLOG (Mar. 16, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.html.

168 Goldsmith & Woods, *supra* note 26.

169 *Id.*

170 *YouTube Community Guidelines Enforcement*, GOOGLE: GOOGLE TRANSPARENCY REPORT, https://transparencyreport.google.com/youtube-policy/removals (last visited Mar. 28, 2020).

171 *What is Project Zero?*, AMAZON, https://brandservices.amazon.com/projectzero (last visited Mar. 28, 2020).

172 Michael Kan, *Facebook Taps Next-Gen AI To Help It Detect Hate Speech*, PCMAG (May 1, 2019), https://www.pcmag.com/news/368104/facebook-taps-next-gen-ai-to-help-it-detect-hate-speech.

large sets of data and teaching the AI to recognize characteristics inside them, the new approach is designed to skip the labeling process and "predict what might be present in the raw training data."[173] Such a new technique was recently implemented by Facebook to block misinformation, especially fake product ads, in relation to the coronavirus.[174]

To make sure the content they host is legitimate, platforms may offer different types of automatic flagging and dispute resolution systems and adjudicate conflicting claims of copyright holders and users, as well as victims of defamatory statements and hate speech.[175] One prominent example is YouTube's Content ID discussed earlier.[176] Instead of resolving a claim of copyright infringement in court—a relatively expensive and time-consuming process—Content ID affords an alternative, private dispute resolution system that allows the parties to settle easily and quickly.[177] As reported, since its launch in 2007,

> Content ID has been updated to use smarter fingerprinting that can detect tricks like stretching a video's aspect ratio, flipping the image horizontally, or slowing down the audio. It has also been plugged into Google's machine learning algorithms. In addition to detecting copyrighted video and audio—thanks to a massive database of over 600 years' worth of reference content provided by networks, record labels, and other rights holders—Content ID can now detect melodies as well.[178]

Another adjudicative/preventive use of AI is for "takedown and stay down" purposes, which involves active monitoring to make sure objectionable content is not re-uploaded.[179] This type of system is based on prediction and prevention.[180] It aims to apply preventive measures by predicting a behavior that has not occurred yet, and possibly may never occur. Another example of a preventive use of AI is the deployment of ML to live chats of users and to the metadata of videos to predict copyright infringement in live video streams.[181]

Finally, AI could be used not only for removing speech, but also for blocking speakers.[182]

---

173 *Id.*

174 Tekla S. Perry, *How Facebook Is Using AI to Fight COVID-19 Misinformation*, IEEE SPECTRUM (May 12, 2020, 4:00 PM), https://spectrum.ieee.org/view-from-the-valley/artificial-intelligence/machine-learning/how-facebook-is-using-ai-to-fight-covid19-misinformation.

175 Kate Crawford & Tarleton Gillespie, *What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint*, NEW MEDIA & SOC'Y 1 (2014), http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.820.1394&rep=rep1&type=pdf.

176 *See supra* notes 135–63 and accompanying text.

177 Sharon Bar-Ziv & Niva Elkin-Koren, *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*, 50 CONN. L. REV. 339, 351–52 (2018).

178 Titlow, *supra* note 162.

179 Annemarie Bridy, *Copyright's Digital Deputies: DMCA-Plus Enforcement by Internet Intermediaries*, *in* RESEARCH HANDBOOK ON ELECTRONIC COMMERCE LAW 185 (John A. Rothchild ed. 2016).

180 ETHAN KATSH & ORNA RABINOVICH-EINY, DIGITAL JUSTICE: TECHNOLOGY AND THE INTERNET OF DISPUTES 52–54, 125–30 (2017).

181 Zhang et al., *supra* note 93, at 369.

182 Jeremy Kahn, *Meet the A.I. that Helped Facebook Remove Billions of Fake Accounts*, FORTUNE

3. Law Enforcement

Platforms also engage in content moderation for law enforcement purposes.[183] Content moderation performed in this capacity could be in compliance with a court order, action responding to a governmental warrant, or otherwise explicitly required by law. Specifically, platforms are facing a growing number of formal and informal requests from law enforcement agents to remove suspicious content. According to Google Transparency Report, during the month of June 2010, Google received 1,181 governmental requests to remove content; in June 2013, this number increased to 3,846 requests; in June 2016, the number of requests was 6,554; and during the month of June 2018, it went up to 25,534 requests.[184]

Additionally, platforms are facing global political and governmental pressure to hone their gatekeeping functions and censor content amounting to hate speech, terrorist propaganda, and pedophilia.[185] An increasing number of recent laws require platforms to act fast and efficiently to remove illicit content. One example is the Singaporean Protection from Online Falsehoods and Manipulation bill from October 2019, which facilitates the blocking of sites promoting fake news pursuant to a governmental order.[186] Another example is the Act to Improve the Enforcement of Rights on Social Networks (NetzDG), which was adopted in Germany in 2017. The law requires intermediaries to delete content which is "clearly illegal" within 24 hours of a complaint being filed.[187] Equivalent initiatives were introduced in the United Kingdom and the Russian Federation.[188] Similarly, a recent proposal by the European Commission would require hosting service providers to remove or disable access to terrorist content within one hour of receipt of a removal order.[189]

Other legal initiatives may indirectly impose removal duties. For instance, calls to abolish the longstanding safe harbor regime,[190] which currently exempts online

---

(Mar. 4, 2020, 4:00 AM), https://fortune.com/2020/03/04/facebook-a-i-fake-accounts-disinformation/.

183 *See generally* Perel & Elkin-Koren, *supra* note 21, at 481.

184 *Government Requests to Remove Content*, GOOGLE: GOOGLE TRANSPARENCY REPORT, https://transparencyreport.google.com/government-removals/overview (last visited July 13, 2020).

185 Elkin-Koren & Perel, *supra* note 90, at 669.

186 *Facebook Expresses 'Deep Concern' After Singapore Orders Page Block*, BBC (Feb. 19, 2020), https://www.bbc.com/news/world-asia-51556620.

187 Beschluss des Bundesrates [Federal Council Decision], Bundesrat Drucksachen [BR] 536/17 (Ger.).

188 *See* Kaesling, *supra* note 110, at 152–56.

189 Press Release, European Commission, State of the Union 2018: Commission Proposes New Rules to Get Terrorist Content off the Web (Sept. 12, 2018), http://europa.eu/rapid/press-release_IP-18-5561_en.htm.

190 *See* Ending Support for Internet Censorship Act, S. 1914, 116th Cong. (2019). The bill, introduced by freshman Sen. Josh Hawley in June, 2019, intends to allow companies to keep their immunity granted under Section 230 of the Communications Decency Act if they submit to an external audit by the Federal Trade Commission that "proves by clear and convincing evidence that their algorithms and content-removal practices are politically neutral." See also Article 17 of Digital Single Market Directive, which holds hosting platforms liable for infringing content posted by their users, unless they have acquired a license or taken measures to prevent the availability of

intermediaries from liability for material hosted by their systems, may induce platforms to undertake preventive measures.[191] In the context of copyright, for instance, the rhetoric is rather straightforward: platforms benefit from the sharing of content, they have the power to efficiently and effectively guard against unwanted content, and, if held liable for users' content, they will act to address the spread of illegal content.[192] In the political context, the Executive Order on Preventing Online Censorship accused online platforms of "engaging in selective censorship that is harming our national discourse."[193] The President's order therefore seeks to narrow the immunity granted to online platforms under Section 230 of the Communications Decency Act, for hosting content generated by their users.[194]

As we explain elsewhere, holding platforms liable for the content they host will likely encourage them to exploit technological filters to screen out illegal content before it ever becomes publicly available.[195]

For instance, Facebook has recently admitted that 99% of the terrorist content they remove is flagged by their AI-based systems before anyone on their services reports it.[196] YouTube has announced that it is using AI to spot extremist content, and that more than 83% of the videos it deleted were flagged by AI, and that three quarters of those were deleted before getting any views.[197] Following the initiative of Tech Against Terrorism, which is supported by the United Nations Counter Terrorism Executive Directorate, several platforms, including Facebook and Microsoft, have been working together to tackle terrorist propaganda by using AI.[198]

E. Multiple Functions and the Public/Private Divide

The discussion above demonstrated the different functions performed by AI content moderation systems, ranging from content matching driven by commercial

---

infringing content from the outset. Council Directive 2019/790, art. 17, 2019 O.J. (L 130) 119–20 (EU). Taking into account global effects, this legislation will inevitably impact the U.S. market as well. *See* Niva Elkin-Koren, Yifat Nahmias, & Maayan Perel, *Is It Time to Abolish Safe Harbor? When Rhetoric Clouds Policy Goals*, 31 STAN. L. & POL'Y 1, 7, 9–11 (2020).

191 Specifically, the safe harbor provisions of the Digital Millennium Copyright Act (DMCA) and § 230 of the Communications Decency Act were intended to strengthen the democratic nature of the internet and promote diversity and participation by facilitating an open and accessible public sphere; *See* Elkin-Koren et al., *supra* note 190, at 7, 9–11.

192 *Id.*

193 Exec. Order No. 13925, Preventing Online Censorship, 85 Fed. Reg. 34079 (May 28, 2020).

194 *Id.* § 2.

195 Elkin-Koren et al., *supra* note 190, at 44–45.

196 Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/.

197 David Meyer, *AI is Now YouTube's Biggest Weapon Against the Spread of Offensive Videos*, Fortune (Apr. 24, 2018, 2:56 AM), https://www.wired.com/2017/04/facebook-live-murder-steve-stephens/; Kate O'Flaherty, *YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks*, WIRED (June 26, 2018), https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video.

198 Keller, *supra* note 5, at 6–7.

interests to a variety of adjudicatory functions. Could these different functions fit neatly within the firm divide between public and private?

Arguably, the art of matching content to users, which is the core of the platform's business model, would not be subject to First Amendment scrutiny but likely be governed by the principles of civil law.[199] Nevertheless, the multiple roles of platforms may challenge this analysis, especially when the structural operation of platforms puts them in a systematic conflict of interests. For instance, assume that a platform is generating more revenues from content A, either because it is its own content or because it generates some income from its business partners. In a private/business capacity, it would be legitimate to give more visibility to such content, but could that be justified in a law enforcement setting where the First Amendment protects users' speech?

A similar challenge is raised in respect to the function of content moderation for law enforcement purposes, which *may* qualify as state action and trigger constitutional scrutiny.[200] As we explained, when platforms remove content in compliance with state warrants, they could be viewed as law enforcement agents and be exposed to claims for violating the First Amendment if they fail to adhere to constitutional free speech standards.

Yet, applying this test to particular cases might be tricky. Platforms may cooperate with law enforcement agents "under the radar," targeting content without any prior formal authorization.[201] In such cases, users might have no clear constitutional means to raise defenses based on the First Amendment.[202] Moreover, governments may require or encourage platforms to put in place their own rules or "Community Guidelines" that will prohibit the promotion of illegal content.[203] In such cases, it is ultimately the platforms themselves who shape and adapt their internal "laws of flagging" pursuant to which the legitimacy of content is determined.[204] While the definition of these internal laws could be informed by firm legal concepts, such as

---

199 Manhattan Cmty. Access Corp. v. Halleck, No. 17-1702, slip op. at 13 (587 U.S. __ June 17, 2019) ("[W]hen a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment because the private entity is not a state actor. The private entity may thus exercise editorial discretion over the speech and speakers in the forum.").

200 Lee, *supra* note 124, at 1057; Perel & Elkin-Koren, *supra* note 21, at 482–83. *But see Halleck*, slip op. at 13 (holding that regulation in and of itself does not transform a private action into a state action).

201 Will Carless & Michael Corey, *Inside Hate Groups on Facebook, Police Officers Trade Racist Memes, Conspiracy Theories and Islamophobia*, REVEALNEWS (June 14, 2019), https://www.revealnews.org/article/inside-hate-groups-on-facebook-police-officers-trade-racist-memes-conspiracy-theories-and-islamophobia/. Edward Snowden's revelations demonstrated that law enforcement agencies were able to retrieve a large volume of data and use it for national security and law enforcement purposes without a warrant, through informal collaboration with digital platforms. Ewen MacAskill & Dominic Rushe, *Snowden Document Reveals Key Role of Companies in NSA Data Collection*, GUARDIAN (Nov. 1, 2013, 5:40 PM), https://www.theguardian.com/world/2013/nov/01/nsa-data-collection-tech-firms; Hanna Kozlowska, *Facebook is Giving the US Government More and More Data*, QUARTZ (Dec. 19, 2017), https://qz.com/1160719/facebooks-transparency-report-the-company-is-giving-the-us-government-more-and-more-data/.

202 Keller, *supra* note 5, at 3–4.

203 *Id.* at 6.

204 *See also* Elkin-Koren et al., *supra* note 190, at 35–36.

"newsworthiness" and "public figures" that have traditionally shaped courts' decisions about defamation,[205] the interpretation and implementation of these concepts by online platforms is ultimately tweaked by their non-transparent, private considerations.[206] Consider, for instance, Article 17 of the recently approved Copyright in the Digital Single Market Directive,[207] which explicitly articulates that an online content-sharing service provider may become directly liable for copyright infringements on the part of its users, unless it has acquired a license or taken measures to prevent the availability of infringing content from the outset. It is fairly anticipated that this legislation will push platforms to deploy AI-based filtering technologies that will screen out allegedly infringing content.[208] Similar consequences may be impelled by Trump's recent Executive Order, if it would actually lead to a narrower application of platforms' immunity under Section 230, pushing social media services to "be far more aggressive in moderating content and terminating accounts."[209] While platforms removing content as the long hand of the government seems like an exercise of public powers, designing an optimal filtering technology involves private, discretionary choices regarding efficiency, accuracy, and cost. Thus, law enforcement by platforms is far more than merely an exercise of public functions.

Even less straightforward is content adjudication between conflicting claims of users and third parties, which may fall in between the public/private distinction. Often these disputes would be based on platforms' community guidelines, reflecting a business choice of risk management, potentially exposing the platform to legal liability towards the parties involved or to commercial sanctions by some communities of users. In some cases, however, adjudication may also impact users' fundamental rights and should therefore invoke constitutional interests. Consider Twitter's recently released fact-checking feature, which labels tweets with potentially misleading or false claims.[210] While preventing the spread of misinformation online definitely promotes the private interest of Twitter in protecting the reliability of its brand, it was also described as the imposition of "unchecked power" to censure the views of others—a description which fairly suits a public actor.

In summary, strictly dividing the functions of content moderation into public

---

205 Kadri & Klonick, *supra* note 21, at 41.

206 *Id.* at 41–42.

207 *Commission Proposal for a Directive of the European Parliament and of the Council on Copyright in the Digital Single Market*, at 16, COM (2016) 593 final (Sept. 14, 2016). The text of the Directive was adopted by the European Parliament on March 26, 2019, with 348 votes in favor, 274 against, 36 abstentions, and 93 MEPs not attending the session. It was subsequently ratified by the European Council. *See* Press Release, European Digital Rights, Censorship Machine Takes Over EU's Internet (Mar. 26, 2019), https://edri.org/censorship-machine-takes-over-eu-internet/; *see also* Martin Husovec, *How Europe Wants to Redefine Global Online Copyright Enforcement*, *in* PLURALISM OR UNIVERSALISM IN INTERNATIONAL COPYRIGHT LAW 529 (Tatiana Eleni Synodinou ed., 2019).

208 Elkin-Koren et al., *supra* note 190, at 44–45.

209 David Smith, *Trump Signs Executive Order to Narrow Protections for Social Media Platforms*, GUARDIAN (May 29, 2020, 8:29 PM) (citing Matt Schruers, the president of the Computer and Communications Industry Association), https://www.theguardian.com/us-news/2020/may/28/donald-trump-social-media-executive-order-twitter.

210 Taylor Hatmaker, *Jack Dorsey Explains why Twitter Fact-Checked Trump's False Voting Claims*, TECHCRUNCH (May 28, 2020, 9:44 PM), https://techcrunch.com/2020/05/27/twitter-vs-trump-fact-checking-dorsey/.

actions and private ones to subject them to appropriate standards of scrutiny is extremely complicated because the implementation of these functions essentially combines a little bit of both types of conduct. Moreover, as we demonstrate in the following Section, the technological architecture of AI systems used by platforms for content moderation practically integrates all functions in a manner that is simply inextricable. Hence, to mitigate the privatization of governmental powers, facilitate oversight, and avoid conflicts of interest, it is necessary to consider a different approach to retain the constitutional divide between private and public.

IV. CONTENT MODERATION BY AI: A SYSTEM PERSPECTIVE

We have seen that the constitutional framework calls for a distinction between private action and public functions (Part II). The discussion so far has also demonstrated that online content moderation by platforms is challenging the public/private constitutional divide (Part III). Still, dividing the different functions performed by social media platforms is necessary for applying different levels of scrutiny and making sure that governmental actions are held to a higher constitutional standard, while ensuring freedom and autonomy of private actors with respect to their private actions. This legal framework assumes that separating between different functions of the same entity is not only desirable but also feasible.

Yet, when content moderation is implemented by AI, the different functions of content moderation are all embedded in a single system, which shares the same data, logic, and learning that shapes the final outcome. The same features of the content matching function, which were designed to maximize profits for the platform, are also applied in performing public functions. Consequently, public functions, which are bound by governmental constraints on speech, might be biased towards commercial interests in non-transparent ways.[211]

Focusing on *how* content moderation is actually performed, the following discussion takes a system approach to describe the technicalities of content moderation by AI. It demonstrates how the different functions of content moderation discussed in Part III converge into a single system. In the next Part, we propose a new approach to facilitate power restraint and accountability in such an integrated system.

A caveat is due here. Very little is publicly known about platforms' content moderation practices, since much of this information is kept secret by platforms behind technical barriers and legal walls of intellectual property. Transparency reports, investigative journalists' stories and occasional leaks by former employees offer a scattered picture of the use of AI in content moderation.[212] The lack of

---

211 Niva Elkin-Koren, *Contesting Algorithms*, BIG DATA & SOC'Y (forthcoming 2020) (manuscript at 12) (on file with authors).

212 Facebook has kept its content moderation guidelines secret for many years. Leaks by a former employee in 2012 offered a first-ever look into Facebook's content moderation system. *See* Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where 'Camel Toes' are More Offensive Than 'Crushed Heads'*, GAWKER (Feb. 16, 2012, 3:45 PM), https://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads; Tarleton Gillespie, *The Dirty Job of Keeping Facebook Clean*, SOC. MEDIA COLLECTIVE (Feb. 22, 2012), https://socialmediacollective.org/2012/02/22/the-dirty-job-of-keeping-facebook-clean/. The Guardian published a subsequent leak of over 100

comprehensive information regarding the actual practices of content removal and the silencing of speakers is part of the accountability crisis in content moderation by platforms.[213] Therefore the following analysis offers a moderately technical description which is based on what has been publicly revealed.[214]

A. How AI is Used in Content Moderation

The deployment of AI tools in content moderation is a sea change in governing speech. The use of ML to identify and remove unwarranted speech is transforming the way laws govern the public sphere.

Content moderation based on ML embeds a dynamic and adaptive decision-making process, which is driven by data.[215] "Supervised learning" is achieved by training the algorithm on previously labeled data (for instance: images labeled "Islamic State propaganda" or labeled "legitimate").[216] Based on sufficient training data, the system will learn to distinguish terrorist propaganda from everything else.[217]

Labeled data could be used to train image-recognition tools that flag unwarranted content. For instance, a system could be trained using on-file matches to identify images of violent uses of guns as similar metadata.[218] Even when an image is not identical, it might be tackled by other types of ML tools such as Digital Hash Technology. Such tools may identify content that is similar, though not identical, to the labeled image. Digital Hash Technology converts images or videos into a hash ("digital signature"), which can be used to identify other iterations of that content.[219] The hash is resistant to alterations, thus enabling the identification of resized images, or images with minor color alterations.[220] This enables the screening of online

documents detailing Facebook's internal content moderation guidelines in 2017. GUARDIAN, *supra* note 108; *see also* Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence*, GUARDIAN (May 21, 2017, 1:00 PM), https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence.

213 Perel & Elkin-Koren, *supra* note 21, at 497; Yemini, *supra* note 21, at 1153.

214 Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money*, HOOVER INST. (Aegis Series Paper No. 1902, 2018), https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf.

215 Press Release, Tech Against Terrorism Awarded Grant by the Government of Canada to Build Terrorist Content Analytics Platform, https://www.techagainstterrorism.org/2019/06/27/press-release-tech-against-terrorism-awarded-grant-by-the-government-of-canada-to-build-terrorist-content-analytics-platform/.

216 Aidan Wilson, *A Brief Introduction To Supervised Learning*, TOWARDS DATA SCI. (Sept. 29, 2019), https://towardsdatascience.com/a-brief-introduction-to-supervised-learning-54a3e3932590.

217 Press Release, Tech Against Terrorism, *supra* note 215.

218 *See* EVAN ENGSTROM & NICK FEAMSTER, THE LIMITS OF FILTERING: A LOOK AT THE FUNCTIONALITY & SHORTCOMINGS OF CONTENT DETECTION TOOLS 11 (2017).

219 Tech Against Terrorism, for instance, has developed a repository of verified terrorist content. *See* Press Release, Tech Against Terrorism, *supra* note 215 (describing "a centralized platform aimed at facilitating tech company moderation of terrorist content and improving quantitative analysis of terrorist use of the internet").

220 PhotoDNA, which was developed by Microsoft, generates hash values of images, video and audio files to identify similar images. *New Technology Fights Child Porn by Tracking Its "PhotoDNA"*, MICROSOFT (Dec. 15, 2009), https://news.microsoft.com/2009/12/15/new-

content, *ex post* or *ex ante*, against a database of predefined illicit content. Every new piece of content identified updates the database and becomes embedded in future screenings by the system.

Unlike rule-based codes, which apply explicit definitions of unwarranted content (e.g., remove x if identical to original content), ML algorithms are deployed to identify patterns and make predictions.[221] Natural Language Processing (NLP) tools, for instance, parsed text in order to make predictions about the sentiment and meaning of the text and to identify hate speech or extremist content.[222]

There are many types of automated tools using ML for content moderation,[223] all sharing similar basic features. Particularly, they all involve datafication (the system choice to collect and record particular data)[224] and the *labeling* of data and its classification as either legitimate or unwarranted. Labeling refers to the recording, aggregating, tagging, and coding of data into a format that could be used for training and data analytics.[225] AI-based content moderation systems further involve a *predictive model*, seeking to predict whether any given content is illicit, based on features learned in the training model, and an automated detection and performance of an *action* (e.g., post, recommend, remove, block, filter). A key feature of ML content moderation systems is a *feedback loop*.[226] Content identified as illicit is fed back into the model so that it will be detected the next time the system runs.[227]

Unsurprisingly, these features of content moderation are also essential to the deployment of ML for tailoring content to users. For instance, Content ID could be used for generating revenues, by allowing right holders to identify their works even when fairly remixed in UGC and share revenues with YouTube. Content ID could also be deployed to remove infringing content or tackle recommendations

---

technology-fights-child-porn-by-tracking-its-photodna/#sm.0001mpmupctevct7pjn11vtwrw6xj. Another example is YouTube Content ID. *YouTube Content ID*, YOUTUBE (Sept. 28, 2010) https://youtu.be/9g2U12SsRns (describing how Content ID creates a recognizable "fingerprint" to identify content).

221 In fact, the great promise of ML lies in addressing "open-ended questions by identifying patterns and making predictions." *See* Jonathan Zittrain, *Intellectual Debt: With Great Power Comes Great Ignorance*, MEDIUM (July 24, 2019), https://blog.usejournal.com/from-technical-debt-to-intellectual-debt-in-ai-e05ac56a502c#0030-637d0839f7b9.

222 Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, NEW AMERICA, 14–15 (July 15, 2019, 10:21 AM), https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/.

223 *Id.* at 5.

224 *See, e.g.*, Katherine J. Strandburg, *Monitoring, Datafication, and Consent: Legal Approaches to Privacy in the Big Data Context*, *in* PRIVACY, BIG DATA AND THE PUBLIC GOOD: FRAMEWORKS FOR ENGAGEMENT 5 (Julia Lane et al. eds., 2015).

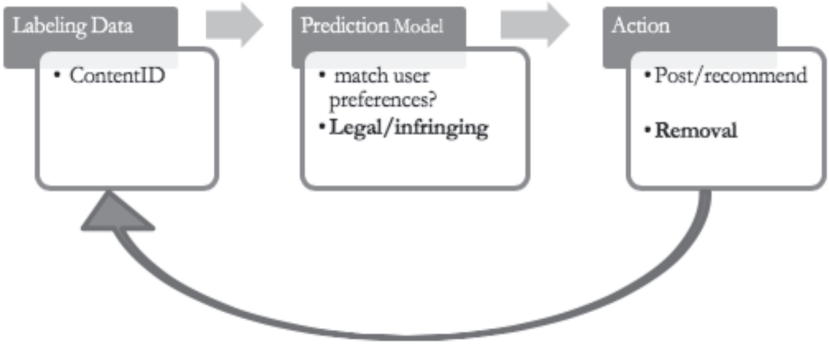
225 Helen Nissenbaum, *Deregulating Collection: Must Privacy Give Way to Use Regulation?* 9 (Cornell Tech, 2017), https://ssrn.com/abstract=3092282 (Data is not simply a raw resource, lying about awaiting collection. Rather, data is "*constructed* or *created* from the signals of countless technical devices and systems.") (emphasis in original).

226 CAMBRIDGE CONSULTANTS, USE OF AI IN ONLINE CONTENT MODERATION 16 (2019) https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

227 *Id.*

made to a particular user.

AI Content Moderation Process



B. Fusion of Functions

As demonstrated above, the different functions of AI content moderation make use of the same data and labeling. In fact, content moderation for law enforcement purposes is built upon the same infrastructure that is designed to personalize content for maximizing profits.

Consequently, both the (private) personalization function of matching users to content and the (public) law enforcement functions are converged in a single system which is informed and shaped by common features. As observed by Jack Balkin: "the infrastructure of free expression is increasingly merging with the infrastructure of speech regulation and the infrastructure of public and private surveillance."[228]

Using a single system for both commercial (private) functions and law enforcement (public) functions may carry some advantages. By making use of existing data and learning acquired by private use, such a system might optimize the public functions performed by platforms to enhance overall efficiency.[229] Yet, the convergence of private and public functions at the system level introduces new legal challenges. The constitutional framework applies different scrutiny to each function, and therefore requires conceptually keeping private and public functions apart. In traditional code environments, different functions would have been performed by discrete programs and therefore would not raise similar challenges.

What makes ML unique is that the system behavior is influenced by the data.[230] Arguably, private functions and public functions are distinct. ML deployed for

---

228 Balkin, *supra* note 3, at 2297.

229 Niva Elkin-Koren & Michal Gal, *The Chilling Effect of Governance-by-Data on Innovation*, 86 U. CHI. L. REV. 403, 405 (2018).

230 D. Sculley et al., *Hidden Technical Debt in Machine Learning Systems*, NIPS, 1 (2015), https://papers.nips.cc/paper/5656-hidden-technical-debt-in-machine-learning-systems.pdf.

recommending content to users is optimizing a commercial goal, while law enforcement systems seek to tackle illicit content. Yet, input from the private system could be consumed by a law enforcement system.[231] The output of any given (commercial) model might be used as an input of another (public) model. The entanglement of private and public functions in ML may thus involve hidden dependencies, which might be difficult to tear apart. D. Sculley and others argue that "[m]achine learning systems mix signals together, entangling them and making isolation of improvements impossible."[232] They call this principle "Changing Anything Changes Everything," arguing that no input is ever really independent and that adding or removing any feature may change the prediction behavior of the system.[233]

Consider for instance the personalization of content to particular users. YouTube, for example, generally seeks to draw a large audience and keep them logged in to generate income from advertising.[234] Thus, YouTube's recommendation engine seeks to show users what they like to watch. Rather than simply using misleading clickbait titles that intend to manipulate users to click the link and view the content, the recommendation system became more sophisticated, applying data analytics to predict user satisfaction.[235] It is designed to hook users to the system by predicting their preferences, based on previous views and recommending content accordingly.[236] YouTube has announced that this system is responsible for about 70% of the time spent by users on its service.[237]

Hence, a user who has watched the NBA championship is more likely to be offered additional sporting events. A user who has searched for information on extreme Islamic ideology might be offered videos on ISIS. Users following up on news items on white supremacists might be offered more videos on radical groups. Thus, the same recommendation engine may lead those who have searched once for some sort of extremist content online to be steered down a radical rabbit hole. Indeed, recent allegations against YouTube claim that the recommendation algorithm "pushes users into . . . a pedophilia 'wormhole'" by "facilitating and monetizing the sexual exploitation of children."[238] The fragmented online discourse may

---

231 Hannah Bloch-Wehba, *Why Police Love the Idea of Automated Content Moderation*, SLATE (Mar. 13, 2020), https://slate.com/technology/2020/03/social-media-content-moderation-surveillance.html.

232 Sculley et al., *supra* note 230, at 2.

233 *Id.*

234 *See supra* Part III.

235 YouTube claims that the ML systems that generate recommendations are trained by using external reviewers using public guidelines. *See External Evaluators and Recommendations*, YOUTUBE HELP, https://support.google.com/youtube/answer/9230586 (last visited Feb. 28, 2020).

236 Roose, *supra* note 155 (describing how YouTube helps radicalize users through its algorithm).

237 Joan E. Solsman, *YouTube's AI is the Puppet Master over Most of What You Watch*, CNET (Jan. 10, 2018), https://www.cnet.com/news/youtube-ces-2018-neal-mohan/ (citing YouTube product chief as saying 70% of the time users watch is driven by "a chain of recommendations run by artificial intelligence").

238 Natasha Lomas, *YouTube Under Fire for Recommending Videos of Kids with Inappropriate Comments*, TECHCRUNCH (Feb. 18, 2019), https://techcrunch.com/2019/02/18/youtube-under-fire-for-recommending-videos-of-kids-with-inappropriate-comments/.

reinforce this: content is displayed to particular users connected to likeminded communities with self-reinforcing power. Since social media does not offer public discourse but rather small tailored "publics," there is less opportunity to be confronted with contesting views.[239]

Platforms were accused of not just allowing hate speech online but also promoting it by the logic of their systems. The alleged harm was caused by datafication, a classification and predictive model that defines another video as "similar" to those previously watched or otherwise catering to the same preferences. YouTube has already considered some steps to reduce the harm caused by its recommendation system, which optimizes the matching of content to users by using the shared features of the system.[240]

Note that the same data, which has been applied to tailor content to particular users, might also be used to address the problem of radicalization by extremist content. This problem could be addressed through datafication, labeling (tagging content as extremist), monitoring the watching habits of those who watch the tagged content, taking actions such as putting viewers on a watch list, and changing the recommendation of content once a user has reached a certain threshold. This type of intervention would be fed back into the system: for instance, presumably, the tagging of content as "extremist" could be shaped by the "type" of users who watch it. Once content is tagged as extremist or even borderline content,[241] it may not be freely shared online.

Sometimes the content itself is not harmful, but the manner in which it has been used could make it harmful to society. The risk arising from social media could be the reinforcement of radical views or conspiracy theories that might lead to violence.[242] Studies have shown that social media platforms could contribute to radicalization due to their feedback loop, which may steer users to increasingly extreme content. For instance, an innocent video of two ten-year-old girls playing in their bathing suits in the neighborhood swimming pool is not harmful content. But such a video could be sexualized if displayed in connection with sexually suggestive videos of women and underage children.[243] A study at Harvard's Berkman Klein Center for Internet and Society found that YouTube's recommendation system curated a list of recommended videos for users that displayed partially clothed children,

---

239 Anat Ben David, *Data in Doubt: Contextualising Facebook Publics in the Age of Political Astroturfing*, NEW MEDIA & SOC'Y 22 (unpublished manuscript) (on file with authors).

240 As YouTube team explains, it is "reducing recommendations of borderline content and content that could misinform users in harmful ways—such as videos promoting a phony miracle cure for a serious illness, claiming the earth is flat, or making blatantly false claims about historic events like 9/11." *See Continuing Our Work to Improve Recommendations on YouTube*, YOUTUBE (Jan. 25, 2019), https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html.

241 Borderline content is not banned by YouTube guidelines and therefore not banned or removed from the system, but instead it is simply not recommended in particular contexts. *See id.*

242 Editorial Board, *The New Radicalization of the Internet*, N.Y. TIMES (Nov. 24, 2018), https://www.nytimes.com/2018/11/24/opinion/sunday/facebook-twitter-terrorism-extremism.html.

243 Max Fisher & Amanda Taub, *On YouTube's Digital Playground, an Open Gate for Pedophiles*, N.Y. TIMES (June 3, 2019), https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html.

sometimes after those users watched sexually explicit content.[244] Each family home video on its own is perfectly innocent but, when grouped together in a particular path of consumption by users following sexually explicit materials, their meaning could change.[245] This case demonstrates how YouTube's ML algorithms may inadvertently become a system promoting pedophilic behavior.

To fix this, YouTube may need to change its recommendation system to exclude videos of children, but this may conflict with its content matching business model. Instead, YouTube has responded with several actions intended to ensure the safety of children,[246] including restricting live features (including new classifiers—ML tools that help identify specific types of content—on their live products to detect and remove live content by minors in violation of this policy),[247] disabling comments on videos featuring minors where users are commenting "inappropriate things,"[248] and limiting recommendations of "borderline content" (now also including videos featuring minors in risky situations).[249]

The bottom line is that in ML, the public function of enforcing legal restrictions (e.g., hate speech, extremist speech, speech in contrary to national security laws) converges with the private functions of platforms. The outcome of optimizing one type of function could shape the outcome of optimizing the other type of function. Lawful content could be excluded from public discourse because it was labeled by the private optimization system as "extremist" (for the purpose of optimizing content matching) and fed into the public optimization system, thereby causing the content to be removed (for the purpose of removing unlawful content). Similarly, content labeled as "borderline" by the public optimization system (because it could be unlawful) could be fed into the private optimization system, which will not only leave it online, but even recommend it to a risky group of extremist users.

---

244 *Id. See generally* Jonas Kaiser & Yasodara Córdova, *On YouTube's Digital Playground*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y AT HARV. U. (June 3, 2019), https://cyber.harvard.edu/story/2019-06/youtubes-digital-playground (discussing YouTube's recommendation algorithm directing users towards harmful content).

245 As explained in the New York Times report: "So a user who watches erotic videos might be recommended videos of women who become conspicuously younger, and then women who pose provocatively in children's clothes. Eventually, some users might be presented with videos of girls as young as 5 or 6 wearing bathing suits, or getting dressed or doing a split." Fisher & Taub, *supra* note 243.

246 *See An Update on Our Efforts to Protect Minors and Families*, YOUTUBE (June 3, 2019), https://youtube.googleblog.com/2019/06/an-update-on-our-efforts-to-protect.html.

247 *Id.*

248 K.G. Orphanides, *On YouTube, A Network of Pedophiles is Hiding in Plain Sight*, WIRED (Feb. 20, 2019), https://www.wired.co.uk/article/youtube-pedophile-videos-advertising (reporting that predators were using the comment sections of YouTube videos with children to guide other pedophiles to the harmful content).

249 "Reducing recommendations: We expanded our efforts from earlier this year around limiting recommendations of borderline content to include videos featuring minors in risky situations. While the content itself does not violate our policies, we recognize the minors could be at risk of online or offline exploitation. We've already applied these changes to tens of millions of videos across YouTube." YOUTUBE, *supra* note 246.

V. SEPARATION OF FUNCTIONS

Decisions regarding the availability of speech and the removal or blocking of expression and speakers are dominated by platforms and executed by AI. This shifts lawmaking power, executive power, and judicial discretion from governmental actors to market players. The constitutional framework seeks to secure basic liberties by relying on the public/private divide. It assumes that governmental powers are attained and exercised by governmental actors, which are subject to the rule of law. When platforms make use of AI systems to execute core public functions, they should ensure that their design and operation comply with constitutional rights. Speech of private actors, however, is protected against unconstitutional restraint by governmental actors.[250] Legal scrutiny demands a distinction between private functions, which allow the platform to exercise editorial discretion, and public functions, which are subject to a higher standard of review.[251]

AI-driven content moderation systems create an integrated fusion of public and private functions in a single system designed to maximize profits for platforms, but at the same time required to perform public functions of law enforcement and judicial judgment. Consequently, AI content moderation systems are incompatible with the constitutional public/private framework intended to secure civil liberties. This creates a new type of democratic deficit as it facilitates the rise of unchecked power, which could escape traditional schemes of checks and balances and constitutional restraints.

Moreover, the difficulty of distinguishing between the flagging of unwarranted content by platforms as private actors and actions performed as state actors also creates a gap in civil enforcement, which could provide another important check over content moderation by platforms. Specifically, the lack of clear distinction between public and private actions could frustrate users' ability to seek remedy under civil law for harm suffered as a result of illegal termination of users or illegal silencing of speech.[252] For instance, Spain's data protection regulator recently held that Google, as a search engine, does "not recognize a legal right of publishers to have their contents indexed and displayed, or displayed in a particular order."[253] Hence,

---

250 *See supra* Part III.D.1.

251 Disclosure duties, transparency reports, and other measures of oversight suffer from major limitations in the context of AI. Content moderation by AI is obviously less transparent than governance by explicit legal norms. The algorithm is opaque. Even if the system's objectives and metrics are explicitly announced, much depends on the implementation of these high level values in the code. Where norms (e.g., features, weight) could be made explicit, the dynamic nature of ML systems could simply make this less useful. Sometimes, as in the case-neutral networks, the process of extracting patterns from data is not even transparent to those who conduct it. Extracting patterns from data might be treated by the data scientists themselves as a "black box," and the link between input data and outcome is often inexplicable. Transparency reports and public oversight have generally proven futile in ensuring these algorithmic regimes advance social welfare.

252 Keller, *supra* note 5, at 2.

253 *See* David Erdos, *Communicating Responsibilities: The Spanish DPA Targets Google's Notification Practices When Delisting Personal Information*, INT'L FORUM FOR RESP. MEDIA: INFORRM'S BLOG (Mar. 21, 2017) (quoting *Guidelines on the Implementation of the Court of Justice of the European Union Judgement on "Google Spain v. Agencia Espanola de Proteccion de Datos (AEPD) and Mario Costeja Gonzalez" C-*

even though Google ranks and removes search results in accordance with its legal obligation under the right to be forgotten, a publisher has no legal right to challenge Google's private decisions about removal and ranking.

In the United States, § 230 of the Communications Decency Act immunizes platforms from most claims based on user content, subject to several statutory exceptions (e.g., intellectual property and criminal liability).[254] By shielding platforms from liability, Congress enabled them to engage in content moderation without risking liability, thereby facilitating more freedom of expression by users. Indeed, this section is considered by many to be the most important driver of online free speech.[255] In a recent Executive Order, President Trump declared a war over its broad application, hoping to constrain the power of platforms to censor allegedly legitimate speech. However, this is unlikely to facilitate a better check over platforms' content moderation practices, but could potentially push them to become more aggressive in silencing borderline content, while remaining unaccountable to their decision-making processes.

A meaningful check over platforms' content moderation could be achieved if platforms are held accountable for the public functions they exercise by subjecting these acts to objective and external review. The charges brought recently by the U.S. Department of Housing and Urban Development against Facebook, alleging its ad-targeting practices discriminate against certain demographics,[256] support this contention.

Aware of the potential risk to the First Amendment in abolishing § 230, some reform proposals seek to amend § 230 to enhance accountability while avoiding a constitutional challenge. For instance, Citron and Wittes propose that § 230 be limited, and not apply when a platform failed to address the illegality of content of

---

131/12, at 10, WP 225 (Nov. 26, 2014)), https://inforrm.org/2017/03/21/communicating-responsibilities-the-spanish-dpa-targets-googles-notification-practices-when-delisting-personal-information-david-erdos/.

254 47 U.S.C § 230. *See generally* Marshall's Locksmith Serv. Inc. v. Google, LLC, 925 F.3d 1263, 1267 (D.C. Cir. 2019) (finding that "Congress[] inten[ded] to confer broad immunity for the re-publication of third-party content"); Doe v. Backpage.com, LLC, 817 F.3d 12, 18 (1st Cir. 2016) ("There has been near-universal agreement that Section 230 should not be construed grudgingly."); Jones v. Dirty World Entm't Recordings, LLC, 755 F.3d 398, 408 (6th Cir. 2014) (quoting Fair Hous. Council of San Fernando v. Roommates.com, LLC, 521 F.3d 1157, 1174 (9th Cir. 2008) (en banc)) ("[C]lose cases . . . must be resolved in favor of immunity."); Doe v. MySpace, Inc., 528 F.3d 413, 418 (5th Cir. 2008) ("Courts have construed the immunity provisions in § 230 broadly in all cases arising from the publication of user-generated content."); Almeida v. Amazon.com, Inc., 456 F.3d 1316, 1321 (11th Cir. 2006) ("The majority of federal circuits have interpreted [Section 230] to establish broad . . . immunity."); Carafano v. Metrosplash.com, Inc., 339 F.3d 1119, 1123 (9th Cir. 2003) (citation omitted) ("§ 230(c) provides broad immunity for publishing content provided primarily by third parties."); Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) ("Congress recognized the threat that tort-based lawsuits pose to freedom of speech in the new and burgeoning Internet medium."); Langvardt, *supra* note 4, at 1369.

255 JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 7–12 (2019); Eric Goldman, *Why Section 230 Is Better than the First Amendment*, 95 NOTRE DAME L. REV. ONLINE 33, 36 (2019) (arguing that § 230 supplements the First Amendment's protections for free speech).

256 Natalie Gagliordi, *Facebook Charged with Violating Fair Housing Act Through Discriminatory Ad Targeting*, ZDNET (Mar. 28, 2019), https://www.zdnet.com/article/facebook-charged-with-violating-fair-housing-act-through-discriminatory-ad-targeting/.

which plaintiffs are complaining.[257] Another initiative proposes to remove § 230 immunity in the case of monetized content.[258] These initiatives assume that public and private aspects of content moderation could be neatly separated. As we have demonstrated, however, these public and private functions are currently converged in online content moderation system.

A. Private Platform–Public Tools

To revive constitutional oversight of the exercise of public powers in content moderation, we propose to translate the public/private divide into a technological idiom, so it can be implemented in an ecosystem governed by AI. The idea is to strengthen civil liberties by introducing *separation of functions* into the system of content moderation to assure public functions (e.g., law enforcement tasks) are kept distinct from private ones and properly scrutinized.

Implementing the *separation of functions* approach in content moderation by AI would involve introducing an independent tool for labeling content and predicting compliance with public standards. The implementation of such a policy could take different forms. One option is to build two separate layers of AI moderation, each independent in its labeling, optimization, and feedback loop. In this case, platforms could be obliged to explicitly separate their internal flagging system (which implement their particular "Community Guidelines" reflecting their standing on free speech issues) from public flagging systems (which implement different statutes, court decisions, or regulatory guidelines defining specific categories of illegal content). The first layer will be internal and optimize the business interests of the platform, effectively conforming to its private content matching interests. This layer will follow the platforms' internal labeling. The second layer will be external and optimize the public interest in removing unlawful content while conforming to law enforcement goals. Both layers will run on the same content hosted and shared on the platform, but they will be completely independent in their labeling, optimization, and feedback loop.

Consider, for example, online copyright enforcement by YouTube. Today, YouTube deploys a single AI-based system, which is converging copyright enforcement (a public function) with negotiating deals between right holders and content providers (a private function). Assume we wish to separate these two functions: Content ID will remain YouTube's business model, bestowing right holders with a fast and easy way to monetize their works, while extracting a share from every deal closed. Nevertheless, YouTube could also be required to install an external screening technology, which will detect copyright infringement according to acceptable legal standards.

PEX is an excellent example of an independent copyright labeling mechanism.[259] PEX "integrates . . . proprietary fingerprinting and indexing technologies

---

257 Danielle Keats Citron and Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans Section 230 Immunity*, 86 FORDHAM L. REV. 401, 414–15 (2017).

258 John Bergmayer, *Even Under Kind Masters: A Proposal to Require that Dominant Platforms Accord Their Users Due Process*, PUBLIC KNOWLEDGE (May 21, 2019), https://www.publicknowledge.org/blog/how-to-go-beyond-section-230-without-crashing-the-internet/.

259 PEX, https://pex.com/ (last visited Mar. 28, 2020).

to identify music and video across all major social platforms."[260] It maps the characteristic components of any given audio or video recording to transform it into a compact coded representation.[261] Although designed as a tool for benefitting right holders and optimizing their earning capacities (and not as a tool for protecting general public interests, such as access to information),[262] this tool demonstrates the importance of enabling alternative labeling by an external independent system. Indeed, Content ID labels the works of specific right holders who had partnered with YouTube for monetization purposes. It does not necessarily label UGC or works by amateurs.[263] As explained in YouTube Help, Content ID is only available to owners of "exclusive rights to a substantial body of original material that is frequently uploaded by the YouTube user community."[264] While this group of right holders is clearly likely to optimize YouTube's share in ad revenues, it does not constitute an inclusive group of the relevant stakeholders. Thus, allowing an external tool to identify which works are protected by copyright is crucial for assuring fair enforcement of copyrights.

Ensuring alternative independent labeling of content is only one way of *separating functions* in content moderation by AI.[265] Alternatively, *separation of functions* could be implemented without separating the platforms' labeling function. Independent tools could also predict an outcome and facilitate an action (e.g., the content upload constitutes copyright infringement and thus must be removed), based on its underling optimization and feedback loop. While Content ID might be designed to optimize YouTube's financial interests by signaling more matches to popular content, an independent tool could be designed to optimize the public interest, for example by balancing copyright protection and fair use, in accordance with the governing legal standards.

In this case, an external, independent tool will develop a predictive model based on the platforms' internal labeling. This could restrict platforms' ability to pollute the feedback loop and tweak their predictive model to match their economic interests. Utopia AI Moderator is a good example of such an independent system.[266] Utopia is a "fully automated" and "real-time moderation tool" that "learns from [past] publishing decisions" that the platforms' human moderators made

---

260 *Id.*

261 *Id.*

262 *Pex: Empowering Rights Owners with Fast Accurate Content Tracking*, GOOGLE CLOUD, https://cloud.google.com/customers/pex/ (last visited May 12, 2020).

263 *See, e.g.*, Danny Fratella, *YouTube Releasing Mini Version of Content ID to Creators with 100k Subscribers*, SOC. BLADE (July 12, 2018), https://socialblade.com/blog/youtube-copyright-match-mini-content-id/ (reporting that a mini-version of Content ID, which allows video creators to detect usage of their content, is available only to creators with 100k subscribers and more); *see Qualifying for Content ID*, YOUTUBE, https://support.google.com/youtube/answer/1311402 (last visited Mar. 28, 2020).

264 *Using Content ID*, YOUTUBE, https://support.google.com/youtube/answer/3244015?hl=en (last visited Mar. 28, 2020).

265 Elkin-Koren, *supra* note 211, at 13.

266 UTOPIA AI MODERATOR, https://utopiaanalytics.com/utopia-ai-moderator/ (last visited Mar. 28, 2020).

previously.[267] The company markets their services as: "your rules, our tools."[268] Hence, for instance, if YouTube is required to use an independent filter like Utopia for copyright enforcement, Utopia will apply its own predictive model on every piece of content identified by Content ID to determine whether it constitute copyright infringement. The system would ban the removal of content found to be lawful. Yet, because we seek to assure the independency of this tool, instead of merely learning from Content ID's internal input, the independent tool would be designed to learn from publicly informed decisions (such as decisions made by judges or classifications by human moderators working for different not-for-profit organizations such as libraries and universities).[269]

To summarize, *separation of functions* in an AI content moderation system involves separating the technical features that predict unlawful content from the platform's private system of content moderation, which is set to optimize its business interests. Rather than reconfiguring the original AI system of content moderation and attempting to alter the optimization model, the public flagging system would be made distinctly separate and independent.

B. Implications and Limitations

The *separation of functions* approach reflects a fundamental principle in administrative law. As Justice Scalia explained, "[s]eparation of functions is a principle of administrative law which seeks to protect the independence and the objectivity of the adjudicative function by restricting its combination with inconsistent functions, such as prosecution, investigation, or advocacy."[270]

As we explained, a technological idiom of this approach could be found in the introduction of an independent framework for flagging unwarranted content. As we have demonstrated, a monolithic system for flagging unwarranted content is dominated by the commercial interests of platforms and their business partners and does not sufficiently safeguard civil rights. *Separation of functions* could revive the public/private divide in an ecosystem where a common good (public discourse) is governed by AI systems that are developed by powerful platforms. *Separation of functions* would distinguish law enforcement functions executed by platforms from their private decisions regarding the removal of content to which they might be held responsible under civil law. It could further subject public conduct to a higher standard of judicial review.

Besides assuring the independence and objectivity of choices about removing unlawful content and safeguarding against conflicts of interests, *separation of powers* would also restrain the concentration of power by platforms.[271] By facilitating the buildup of independent AI capabilities, either by government or by third parties, *separation of powers* will limit platforms' current dominance over the shape of our public discourse. Lawful content unjustifiably flagged as "unlawful" will stop

---

267 *Id.*

268 *Id.*

269 Elkin-Koren, *supra* note 211, at 19.

270 Antonin Scalia, *Separation of Functions: Obscurity Preserved*, 34 ADMIN. L. REV. v (1982).

271 Robert B. Reich, *Big Tech Has Become Way Too Powerful*, N.Y. TIMES (Sept. 18, 2015), https://www.nytimes.com/2015/09/20/opinion/is-big-tech-too-powerful-ask-google.html.

evaporating from our online public sphere with no reason; unlawful content will be strictly removed, even if its pervasiveness is worth a fortune. Platforms will retain their discretion over content matching, but at the same time held accountable to the exercise of law enforcement powers.

Separation of functions should also promote innovation. Platforms are often blamed for locking their data, while creating barriers for ML.[272] By putting a wall around their data, platforms are not only limiting the public ability to scrutinize their governing functions but are also hindering the development of alternative content moderation systems. Allowing independent systems to run on the platforms' data would provide the indispensable resource necessary for developing alternative competing systems of content moderation by AI. This could create a market of independent systems for identifying legitimate content in additional contexts. A competitive market in AI for identifying legitimate content may also promote innovation by creating competitive pressures on market players to invest in improving their systems.

Nevertheless, the *separation of functions* approach may suffer from several limitations too. A potential risk of introducing an independent public enforcement function into AI-based content moderation systems is that it may come at the cost of reducing accuracy, efficiency, and effectiveness in the public enforcement function.[273] Specifically, some information obtained by platforms on speech and speakers might be useful for more efficient enforcement, but it may be unavailable to the independent system designed for enforcement purposes, unless authorized by law. The virality of content, for instance, is an important factor that could arguably hone enforcement practices, as it may be a proxy for radicalism.[274] Moreover, data about *who* consumes the content may also assist in flagging it correctly. Indeed, there are various strategies to hide the illegality of content.[275] Data about specific users who consume apparently innocent content may reveal critical insights about the content's actual meaning. For instance, metadata monitoring of content, which is essentially designed to search the file's metadata or other textual tags attached to it to match it to an existing catalog of files, could be easily circumvented by users.[276] Instagram, for example, has been removing posts containing nudity using hashtags that suggest nudity (e.g., #boobs, #naked). However, users had soon manipulated the platform's algorithm by using umlauts, cedillas, and accents (e.g., #pörnöö).[277] If Instagram could inspect whether these users have consumed unwarranted nudity in the past, it could arguably gain an important advantage in enforcing its policies in the future. Similar concerns were often raised in administrative contexts where

---

272 *Id.*

273 *See* Sculley et al., *supra* note 230.

274 *See supra* Part IV.B.

275 One famous example are the ways users used to signal Jewish names in the Google Chrome browser to subject them to anti-Semitic abuse. *See Google Bans Plug-In That Picks Out Jews*, BBC NEWS (June 6, 2016), https://www.bbc.com/news/technology-36459990.

276 Metro-Goldwyn-Mayer Studios, Inc. v. Grokster, Ltd., 518 F. Supp. 2d 1197, 1206 (C.D. Cal. 2007).

277 Melanie Ehrenkranz, *The Best NSFW Instagram Hashtags Use Special Characters to Hide Porn*, MIC (July 14, 2016), https://mic.com/articles/148675/the-best-nsfw-instagram-hashtags-use-special-characters-to-hide-porn-enjoy#.1Drt7g5Ua.

agencies were forced to strictly separate their functions. Yet, what might be viewed as a "bug" from a system efficiency perspective could actually be conceived as a positive feature from a human rights perspective, encouraging developers to improve accuracy, but not at the cost of compromising privacy.

Another limitation of *separation of functions* relates to the way it might crunch upon platforms' proprietary interests. While the independent tool we proposed would run on the platforms' data and possibly also on its labeled data, it would be installed by platforms and implemented within their "facilities." That is, the idea is not to transfer data from the platforms to an external system, but instead to require platforms to install an external AI mechanism above the private layer of content moderation. Nevertheless, an open question remains: who will own the "learning" acquired by the tool? After all, it is mainly for the massive trove of data and immense volume of online content on prominent platforms such as Facebook and YouTube that the independent tool would be able to prosper. Deciding who should own the rights to the value added to the independent developer is a difficult question that mixes policy concerns with market considerations. We will leave it for another day.

Finally, *separation of functions* is not expected to work on a voluntary basis. We anticipate that platforms would be deterred to install external AI tools, especially if such installment is costly and likely to negatively affect their private functions. Indeed, anecdotal evidence indicates that Facebook has previously turned down offers to use Utopia AI analytics for content moderation purposes.[278] Similarly, if YouTube installed a PEX-like technology for copyright enforcement it could risk losing its ability to monetize fair uses of copyrighted content. Accordingly, the promise of *separation of functions* depends on mandating its implementation through proper regulation.

VI. CONCLUSION

There is a growing concern among Internet law scholars that the First Amendment is becoming obsolete. Driven by the threat of government censorship to free speech, the constitutional framework seems ill-equipped to constrain speech regulation by powerful online platforms. This paper shows that the gap between First Amendment jurisprudence and online speech regulation is not founded merely on doctrinal disparity, but also, and even more so, on practical limitations. Speech regulation by platforms is now a systematic fusion of private content matching functions and regulatory content moderation. With the increasing use of AI by platforms for content moderation, it is insufficient to try to classify specific functions of content moderation as state actions to facilitate traditional constitutional restraint. This is because the same technical design that platforms deploy to curate personalized content is also applied to monitor and censor online speech. Platforms rely on a single, inextricable system of AI to both maximize their earnings capacity and to minimize their legal risk of liability for unlawful content posted by their users. In this technical ecosystem, traditional public/private classifications largely lose their

278 Press Association, *Facebook 'Turned Down' AI Tool to Stop Hate Speech*, IRISH EXAMINER (June 18, 2019), https://www.irishexaminer.com/breakingnews/world/facebook-turned-down-ai-tool-to-stop-hate-speech-931421.html.

distinctive power. Accordingly, this paper advanced a design-based approach, namely "separation of functions." To separate governmental functions in an AI-driven private system and subject them to a higher standard of judicial review, it is necessary to separate the platforms' data collection and labeling from the technical tools which are designed to perform governmental functions based on that data. Otherwise, data labeling and predictive models executed for private, financial goals will continue to distort the feedback loop of public policy optimization algorithms, and vice versa. By building independent tools that embed public values, a practical separation between independent public tools and private data could be achieved. This functional separation of functions may enhance public scrutiny of speech regulation and also facilitate competition among different players who may enrich the design of speech regulation and mitigate biases.