Automation in Moderation

Hannah Bloch-Wehba[†]

> What is an AutoModerator?
>
> An AutoModerator is a bot designed to automate various moderation tasks that require little or no human judgement. It can watch the new/spam/comments/report queues of any subreddit it moderates and take actions on submissions and comments based on defined conditions. This includes approving or removing them . . . . It is effectively fairly similar to reddit's built-in spam-filter, but [also] allows for conditions to be defined specifically instead of just giving vague hints by removing/approving. Its decisions can

[†] Assistant Professor of Law, Drexel University Thomas R. Kline School of Law; Affiliated Fellow, Yale Law School Information Society Project. I am very grateful to RonNell Andersen Jones, Rebecca Crootof, James Grimmelmann, Anil Kalhan, Jonathan Marks, Christopher Reed, Daniel Susser, and Ari Ezra Waldman for their helpful comments and feedback on this project. Portions of this Article were presented at the *Cornell International Law Journal*'s Symposium, the Penn State Law School Works-in-Progress Workshop, and Yale Law School's Information Society Project. My thanks to the student organizers of the symposium and the editors of the *Cornell International Law Journal.* This Article reflects developments through December 2019, when it was finalized for publication. All errors are my own.

53 Cornell Int'l L.J. 41 (2020)

always be overridden by human mod[erators], exactly like an existing filter.[1]

Introduction

In March 2019, a shooter posted a white nationalist manifesto on "8chan," an online message board, and then livestreamed on Facebook as he murdered fifty-one people at two mosques in Christchurch, New Zealand.[2] In the aftermath of the shooting, millions of users viewed the video on YouTube and Facebook even as the sites struggled to keep the video offline.[3] In July 2019, Brandon Clark murdered Bianca Devins and posted grisly pictures of her corpse on the social media platforms Instagram and Discord before attempting suicide and being arrested.[4] The photos, tagged with the hashtag #RIPBianca, quickly spread throughout social media platforms even as users flagged them and called the police in real time.[5] On Yom Kippur of October 2019, another shooter live-streamed on Twitch, a gaming platform, as he murdered two people in a synagogue in Halle, Germany.[6]

In the wake of these incidents, lawmakers around the world are closely scrutinizing "content moderation"—the set of practices that online platforms use to screen, rank, filter, and block user-generated content. One particularly notable regulatory strategy encourages platforms to use technology to prevent the dissemination of unlawful online content before it is

[1] *What is AutoModerator?*, Reddit (2012), http://www.reddit.com/r/AutoModerator/comments/q11pu/what_is_automoderator [https://perma.cc/2XRR-5525].

[2] *See* Kevin Roose, *A Mass Murder of, and for, the Internet*, N.Y. Times (Mar. 15, 2019) https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html [https://perma.cc/XP8P-MCZR]. *See also* Richard Pérez-Peña, *Two New Zealand Mosques, a Hate-Filled Massacre Designed for Its Time*, N.Y. Times (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/world/australia/new-zealand-mosque-shooting.html [https://perma.cc/B7AK-SC3L].

[3] Elizabeth Dwoskin & Craig Timberg, *Inside YouTube's Struggles to Shut Down Video of the New Zealand Shooting– and the Humans Who Outsmarted Its Systems*, Wash. Post (Mar. 18, 2019, 6:00 AM), https://www.washingtonpost.com/technology/2019/03/18/inside-youtubes-struggles-shut-down-video-new-zealand-shooting-humans-who-outsmarted-its-systems/ [https://perma.cc/7HT9-7MNV]; Kate Klonick, *Inside the Team at Facebook That Dealt with the Christchurch Shooting*, New Yorker (Apr. 25, 2019), https://www.newyorker.com/news/news-desk/inside-the-team-at-facebook-that-dealt-with-the-christchurch-shooting [https://perma.cc/UV28-2RKF] [hereinafter *Inside the Team at Facebook That Dealt with the Christchurch Shooting*]; Cade Metz & Adam Satariano, *Facebook Restricts Live Streaming After New Zealand Shooting*, N.Y. Times (May 14, 2019), https://www.nytimes.com/2019/05/14/technology/facebook-live-violent-content.html [https://perma.cc/T2MH-5BPL]; Charlie Warzel, *The New Zealand Massacre Was Made to Go Viral*, N.Y. Times (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/opinion/new-zealand-shooting.html [https://perma.cc/2YYT-M825].

[4] Michael Gold, *#RIPBianca: How a Teenager's Brutal Murder Ended Up on Instagram*, N.Y. Times (July 15, 2019), https://www.nytimes.com/2019/07/15/nyregion/bianca-devins-death.html [https://perma.cc/44GK-7XLG].

[5] *Id.*; Taylor Romine, *Brandon Clark, Accused of Killing Internet Personality Bianca Devins, Pleads Not Guilty*, CNN (July 29, 2019, 4:38 PM), https://www.cnn.com/2019/07/29/us/bianca-devins-murder-new-york/index.html [https://perma.cc/88E7-CJCQ].

[6] Melissa Eddy et al., *Assailant Live-Streamed Attempt Attack on German Synagogue*, N.Y. Times (Oct. 9, 2019), https://www.nytimes.com/2019/10/09/world/europe/germany-shooting-halle-synagogue.html [https://perma.cc/Q3R5-QEZE].

ever seen or distributed.[7] This Article outlines recent efforts to compel or encourage platforms to engage in automated, *ex ante* monitoring, filtering, and blocking of online content across a variety of contexts—defamation, copyright infringement, and terrorist speech. Proponents of these initiatives suggest that *ex ante* screening requirements will incentivize platforms to promote healthier online discourse. Supporters have also suggested that new efforts to regulate platforms' "content moderation" practices limit Big Tech's power by requiring platforms to bear an appropriate amount of responsibility.[8]

But this new breed of regulation comes with unappreciated costs for civil liberties and unexpected boons for platform power.[9] The new automation techniques exacerbate existing risks to free speech and user privacy, and create new sources of information that can be exploited for surveillance, raising concerns about free association, religious freedoms, and racial profiling. Moreover, the automation process worsens transparency and accountability deficits. Far from curtailing private power, the new regulations expand platform authority to include policing online speech, with little oversight and few countervailing checks. This embrace of automation in moderation displays unwarranted optimism about technology's ability to solve what is fundamentally a social and political problem.

Technology platforms' role as "central players" in governing online speech and surveillance has been the subject of rich and growing scholarly literature.[10] In comparison, the role of automation in this context has

[7] But automated systems currently in place do not always work and platforms often depend on users and third parties to report harmful content. These calls are not limited to ordinary posts by individual users. In spring 2019, Facebook pulled Donald Trump's campaign ads after their automated system failed to detect that the ads violated the platform's guidelines by explicitly targeting voters based on gender. Owen Daugherty, *Facebook Pulls Trump Campaign Ad Violating Platform's Policy*, HILL (Aug. 20, 2019, 2:12 PM), https://thehill.com/policy/technology/458116-facebook-pulls-trump-campaign-ad-violating-platforms-policy [https://perma.cc/L6HR-J2ER]; Judd Legum, *Facebook Admits Trump Campaign Is Violating Its Rules, Takes Down Numerous Ads Targeting Women*, POPULAR INFO. (Aug. 19, 2019), https://popular.info/p/facebook-admits-trump-campaign-is [https://perma.cc/HC2G-85LR].

[8] *See, e.g.*, Kevin Madigan, *Will the EU Finally Hold Internet Giants Accountable?*, CPIP (July 3, 2018), https://cpip.gmu.edu/2018/07/03/will-the-eu-finally-hold-internet-giants-accountable/ [https://perma.cc/SJ9Y-LTXF] (describing online platforms as "the most powerful and wealthy entities in the world"); *Europe Takes an Important Step Toward Platform Accountability with Directive on Copyright*, AAP (Sept. 13, 2018), https://newsroom.publishers.org/europe-takes-an-important-step-toward-platform-accountability-with-directive-on-copyright/ [https://perma.cc/EW35-TNPF]. *See also* Press Release, U.S. Senator Josh Hawley, Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies (June 19, 2019) (available at https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies [https://perma.cc/FMM5-9PSJ]).

[9] This analysis is necessarily preliminary, both because the regulatory landscape has shifted dramatically in the last year, and because technological change is underway. *See infra* Part II (discussing recent regulatory developments).

[10] *See generally, e.g.*, JULIE E. COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM (Oxford Univ. Press 2019); Jack Balkin, Essay, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011 (2018) [hereinafter *Free Speech Is a*

received scant scholarly attention. This article aims to fill that gap by exploring the utility of automation,[11] and how its use in law enforcement may lead to cooptation by powerful actors.[12] Automation affords a new and attractive menu of options for private stakeholders, law enforcement, and intelligence agencies.[13] Automation in content moderation is part of a much broader push for private industry to develop swifter, more accurate, and more effective technologies to aid law enforcement.

Private companies, not state actors, largely control the infrastructure of free speech today.[14] The largely hands-off approach to regulating online intermediaries has also allowed them to develop extraordinary expertise regarding controlling the delivery of online content—harvesting, compiling, and profiting off of vast amounts of user data in the process. Today, the

Triangle]; Anupam Chander, *Facebookistan*, 90 N.C. L. Rev. 1807 (2012); Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 Notre Dame L. Rev. 1035 (2018) [hereinafter *Extremist Speech*]; Jennifer Daskal, *Speech Across Borders*, 105 Va. L. Rev. 1605 (2019); Kristen E. Eichensehr, *Digital Switzerlands*, 167 U. Pa. L. Rev. 665 (2019); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2018); Molly K. Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, 60 Va. J. Int'l L. (forthcoming July 2020); Kyle Langvardt, *Regulating Online Content Moderation*, 106 Geo. L.J. 1353 (2018); Frank Pasquale, *Beyond Innovation and Competition: The Need for Qualified Transparency in Internet Intermediaries*, 104 Nw. U. L. Rev. 105 (2010); Alan Z. Rozenshtein, *Surveillance Intermediaries*, 70 Stan. L. Rev. 99 (2018). *See also* Tarleton Gillespie, Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions That Shape Social Media 177 (Yale Univ. Press 2018); Sarah T. Roberts, Behind the Screen: Content Moderation in the Shadows of Social Media 33 (Yale Univ. Press 2019).

[11] Except for algorithmic copyright enforcement, which has been the subject of sustained examination, automation in moderation has largely escaped scrutiny. *See generally* Annemarie Bridy, *Is Online Copyright Enforcement Scalable?*, 13 Vand. J. Ent. & Tech. L. 695 (2011) [hereinafter *Is Online Copyright Enforcement Scalable?*]; Lital Helman & Gideon Parchomovsky, *The Best Available Technology Standard*, 111 Colum. L. Rev. 1194 (2011); Maayan Perel & Niva Elkin-Koren, *Accountability in Algorithmic Copyright Enforcement*, 19 Stan. Tech. L. Rev. 473 (2016). Many scholars are considering the automation of decision-making in other adjudicatory settings. *See* Danielle Keats Citron, *Technological Due Process*, 85 Wash. U. L. Rev. 1249, 1254 (2008) [hereinafter *Technological Due Process*]; Cary Coglianese & David Lehr, *Regulating by Robot: Administrative Decision Making in the Machine-Learning Era*, 105 Geo. L.J. 1147, 1152 (2017); Cary Coglianese & David Lehr, *Transparency and Algorithmic Governance*, 71 Admin. L. Rev. 1, 1 (2019) [hereinafter *Transparency and Algorithmic Governance*]. *See, e.g.*, Kiel Brennan-Marquez & Stephen E. Henderson, *Artificial Intelligence and Role-Reversible Judgment*, 109 J. Crim. L. & Criminology 137, 137 (2019); Rebecca Crootof, *"Cyborg Justice" and the Risk of Technological-Legal Lock-In*, 119 Colum. L. Rev. F. 233, 233 (2019); Sandra G. Mayson, *Bias in, Bias Out*, 128 Yale L.J. 2218, 2218 (2019); Frank Pasquale, *A Rule of Persons, Not Machines: The Limits of Legal Automation*, 87 Geo. Wash. L. Rev. 1, 2 (2019); Eugene Volokh, *Chief Justice Robots*, 68 Duke L.J. 1135, 1135 (2019).

[12] *See* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 Harv. L. Rev. 2296, 2324-29 (2014) [hereinafter *Old-School/New-School Speech Regulation*].

[13] *See* Shoshana Zuboff, The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (Pub. Affairs 1st ed. 2019).

[14] The prevalence of private ownership has shifted the legal landscape from a dualist system in which states regulate speakers directly, to a pluralist model in which the Internet infrastructure serves as a critical intermediary between states and speakers. *Free Speech Is a Triangle*, *supra* note 10, at 2021. *See also* Klonick, *supra* note 10, at 1617.

private sector's capacity to structure, censor, and control the flow of information far outstrips that of the government.

This reality has given rise to a new regulatory approach. Under the new breed of regulation—typified by initiatives like the European Union's Copyright Directive, Germany's Network Enforcement Act of 2018, and Australia's Abhorrent Violent Material (AVM) statute—platforms must take down unlawful content faster, sometimes within twenty-four hours.[15] Yet, the state does not directly require online platforms to adopt specific methods or techniques to achieve this goal, nor to directly control the outcomes of content moderation decisions. Rather, this new approach imposes demanding obligations on platforms while at the same time yielding to them substantial discretion and enforcement authority, leaving it to the private sector to determine how to comply.[16]

This approach might seem like an appropriate middle ground between command-and-control regulation on the one hand, and self-regulation on the other. Indeed, what might variously be called co-regulation, collaborative governance, or multi-stakeholder governance is an increasingly popular framework for governing the technology sector in multiple contexts far beyond content regulation.[17]

In the context of automated content regulation, however, this approach has several major drawbacks. First, in the absence of clear obligations, platforms will tend to over-censor and over-block. Both, state actors and the private sector, have acknowledged that automated content moderation is both over- and under-inclusive.[18] Automation also creates new sources of information that will be valuable to both, private and public sector actors, and opens the door for further "relational" surveillance of users and their broader networks.[19]

Second, this regulatory paradigm extends law enforcement's influence to the design, process, and substance of automated content moderation. Politics already affect the design and the implementation of content moderation rules like the types of user-generated content that platforms opt to control and the ways in which platforms police that content.[20] Though pri-

[15] *See* discussion *infra* Part II.

[16] *Id.*

[17] *See* Robert Gorwa et al., *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, Big Data & Soc'y, Jan.-June 2020, at 1, 1-2 (describing a movement toward co-regulation and transnational standards); Margot E. Kaminski, *Binary Governance: Lessons from the GDPR's Approach to Algorithmic Accountability*, 92 S. Cal. L. Rev. 1529, 1533 (describing how algorithmic accountability requires both collaborative governance and an individual rights regime).

[18] *See* Gorwa et al., *supra* note 17, at 7-10.

[19] Katherine J. Strandburg, *Freedom of Association in a Networked World: First Amendment Regulation of Relational Surveillance*, 49 B.C. L. Rev. 741, 751 (2008).

[20] *See, e.g.*, Joseph Cox & Jason Koebler, *Why Won't Twitter Treat White Supremacy Like ISIS? Because It Would Mean Banning Some Republican Politicians Too*, Vice (Apr. 25, 2019, 12:21 PM), https://motherboard.vice.com/en_us/article/a3xgq5/why-wont-twitter-treat-white-supremacy-like-isis-because-it-would-mean-banning-some-republican-politicians-too [https://perma.cc/J3AP-HXKH]; Casey Newton, *Why Twitter Has Been Slow to Ban White Nationalists*, Verge (Apr. 26, 2019, 6:00 AM), https://www.theverge.com/

vately developed and implemented, these frameworks are neither apolitical nor neutral.[21] Rather, content moderation rules—and the technologies that apply them—reflect corporate, social, and legal values.[22] Platforms adapt their content moderation rules and practices to conform to regulators' preferences, both to comply and to avoid new regulations.

Perhaps the most significant danger of this approach is that, by designing new technologies of content moderation, platforms will create irresistible tools for law enforcement. Public-private cooperation is at the core of ongoing efforts to fight cybercrime, and investigative methodologies are increasingly rooted in proprietary technology.[23] Although opponents of "Big Tech" often describe automation-in-moderation requirements as a method of checking platform power, many of these new initiatives are more likely to entrench the power of online platforms by making them indispensable to government regulators. In their current form, regulations that demand that platforms build and deploy proactive monitoring and filtering mechanisms, risk aggrandizing the corporate power they ostensibly seek to limit—they entrust the private sector to design its own compliance tools.[24] Moreover, preserving the centralization and dominance of large technology companies is likely to make surveillance cheaper and easier for law enforcement.[25]

interface/2019/4/26/18516997/why-doesnt-twitter-ban-nazis-white-nationalism, [https://perma.cc/E24Z-U4WQ].

[21] *See Old-School/New-School Speech Regulation*, *supra* note 12, at 2298 ("Because there are so many speakers, who are often anonymous, difficult to co-opt, or otherwise beyond the government's effective control, the state aims at Internet intermediaries and other owners of digital infrastructure . . . ."); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 27 (2019); *Technological Due Process*, *supra* note 11, at 1037.

[22] *See* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION 61 (Harvard Univ. Press 2015) ("Despite their claims of objectivity and neutrality, they are constantly making value-laden, controversial decisions."). *See also* Ifeoma Ajunwa, *The Paradox of Automation as Anti-Bias Intervention*, 41 CARDOZO L. REV. 1, 3, 5 (forthcoming 2020); Klonick, *supra* note 10, at 1616 (citing other scholars in support of the argument that Facebook's legal culture is distinctively imbued with American free speech thinking); Harry Surden, *Values Embedded in Legal Artificial Intelligence* 1 (Univ. of Colo. Law Legal Studies, Research Paper No. 17-17, 2017) ("Technological systems can have values embedded in their design."). *See also* Bruno Latour, *Technology Is Society Made Durable*, 38 SOC. REV. 103, 130 (1990); Ari Ezra Waldman, *Power, Process, and Automated Decision-Making*, 88 FORDHAM L. REV. 1, 4 (2019) ("[A]lgorithmic decision-making hides the fact that engineers and their corporate employers are choosing winners and losers while steadfastly remaining agnostic about the social, political, and economic consequences of their work.").

[23] ANDREW GUTHRIE FERGUSON, THE RISE OF BIG DATA POLICING: SURVEILLANCE, RACE, AND THE FUTURE OF LAW ENFORCEMENT 129–30 (N.Y. Univ. Press 2017).

[24] *See* COHEN, *supra* note 10, at 122. ("As network intermediaries have resisted efforts to write the logic of the exception into law, they have become masters at both public relations and inside-the-Beltway political positioning. The result is a legal and media landscape characterized by complex power struggles among the dominant interests. In those struggles, platforms do not simply play defense. Rather, they have worked to position themselves as both essential partners and competing sovereigns in the quest to instantiate states of exception algorithmically.").

[25] Tyler Cowen, *Breaking Up Facebook Would Be a Big Mistake*, SLATE (June 13, 2019, 7:30 AM), https://slate.com/technology/2019/06/facebook-big-tech-antitrust-breakup-

This new breed of content regulation is worthy of its own analysis because it illustrates the "embeddedness" of platforms in politics and the ease with which states can influence ostensibly private regulation in order to censor and surveil.[26] Seen through this lens, the turn toward automation is neither a check on the power of technology companies nor a guarantee that they will act more effectively or more neutrally. Rather, the increasing reliance on automation will heighten the risk that both, platforms and governments, will experience cooptation and capture.[27]

Over-reliance on the private development of new technologies of moderation is thus poor public policy on at least two levels. The turn toward automation poses straightforward, significant risks to user speech and privacy, and fails to encompass meaningful checks against those risks. But in a more political sense, the new regulations disguise themselves as accountability measures while providing the private sector with a source of power and profit and entrenching frameworks through which they are likely to be coopted.

The remainder of the discussion proceeds in four parts. Part I describes how the framework of intermediary immunity permitted large online platforms to experiment with automated moderation technologies, and traces how this experimentation came to characterize modern online platforms. Part II maps several recent legal developments that urge platforms to adopt automated and proactive filtering and monitoring techniques in sectors as far-flung as copyright enforcement, defamation, and violent content. In Part III, the Article explores the normative consequences of these developments, considering how they might aggravate existing tendencies toward censorship and surveillance, encode bias and harmful stereotypes, and aggrandize corporate power. Part IV offers an agenda for moderating the use of automation. Rigorous notice requirements, transparency rules, and independent oversight bodies—elements

mistake.html [https://perma.cc/L8GH-CFBL] ("We're probably better off having major, well-capitalized companies as guardians and gatekeepers of online channels, however imperfect their records, as the relevant alternatives would probably be less able to fend off abuse of their platforms and thus we would all fare worse."). *See also* Jon Bateman, *The Antitrust Threat to National Security*, Wall Street J. (Oct. 22, 2019, 6:43 PM), https://www.wsj.com/articles/the-antitrust-threat-to-national-security-11571784197 [https://perma.cc/CK7S-8J3Q]; Cory Doctorow, *Regulating Big Tech Makes Them Stronger, so They Need Competition Instead*, Economist (June 6, 2019), http://www.economist.com/open-future/2019/06/06/regulating-big-tech-makes-them-stronger-so-they-need-competition-instead [https://perma.cc/E6KY-6J9H].

[26] Sarah T. Roberts, *Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation*, First Monday (Mar. 2018), https://www.firstmonday.org/ojs/index.php/fm/article/view/8283 [https://perma.cc/79XP-2EQC] (describing "the platforms' own 'embeddedness' with the U.S. political establishment, and their own relationship to policy, foreign and domestic."); *Old-School/New-School Speech Regulation*, *supra* note 12, at 2325 ("[T]he government offers a combination of carrots and sticks, the most important being legal immunity for assisting the government in identifying or shutting down Internet sites and speakers that the government disfavors or seeks to regulate.").

[27] *Old-School/New-School Speech Regulation*, *supra* note 12, at 2325–26.

notably lacking from the current initiatives—might promote accountability for both, platforms and state actors.

I. The Origins of Automation in Moderation

As James Grimmelmann defines it, moderation is "the governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse."[28] By "content moderation," I mean a platform's internal decision-making on whether user-generated content violates its rules, and if so, what the penalty might be.[29] By enforcing what Sarah Roberts calls the "rules of engagement in online social spaces," content moderators attempt to delineate between acceptable and unacceptable conduct as the platform defines it.[30]

These functions are inextricably linked to the formation and governance of online communities, but moderation also mitigates the risk that unwanted content might alienate users and reduce platform profits. Increasingly, moderation rules also address the risk that political actors might regulate platforms in ways that would diminish the power of said platforms.[31] "Moderation," thus, has two functions: to constitute rules and procedures for a community, and to limit the "intensity or extremeness" of its substance.[32] By defining the boundaries of participation in a community and imposing sanctions on those who violate the conditions of that membership, moderation rules are at the core of online communities' ability to regulate themselves and shape the conditions for free expression.[33]

This Part begins by explaining how and why intermediary liability laws in the United States (U.S.) and in Europe have historically granted broad deference to platforms' rules and mechanisms for governing user speech. As a result of this deference, platforms were able to develop rules and technologies for blocking, filtering, and monitoring user speech on a voluntary basis.

A. Immunity, Safe Harbor & Private Governance

Because both the U.S. and Europe have observed protections against intermediary liability that formally deferred to self-regulation by online actors, the private governance of online speech is of particular impor-

[28] James Grimmelmann, *The Virtues of Moderation*, 17 Yale J.L. & Tech. 42, 47 (2015) [hereinafter *The Virtues of Moderation*] (emphasis omitted).

[29] Roberts, *supra* note 10, at 1.

[30] *Id.* at 33-35.

[31] *See, e.g.*, James Grimmelmann, *The Platform Is the Message*, 2 Geo. L. Tech. Rev. 217, 217 (2018) [hereinafter *The Platform Is the Message*]; Klonick, *supra* note 10, at 1667-68. *See also* Julian Dibbell, My Tiny Life: Crime and Passion in a Virtual World 20 (Henry Holt & Co. 1st ed. 1998) (considering the "death penalty").

[32] *Moderate*, Merriam-Webster, https://www.merriam-webster.com/dictionary/moderation [https://perma.cc/L9RF-XZWS] (last visited June 23, 2020) ("1: to lessen the intensity or extremeness of; 2: to preside over or act as chairman of . . . .").

[33] *The Virtues of Moderation*, *supra* note 28, at 48-50; *The Platform Is the Message*, *supra* note 31, at 224.

tance.[34] Intermediary protections found their strongest expression in Section 230 of the Communications Decency Act of 1996 (CDA), which states: "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[35] In the CDA's "Good Samaritan" provision, Congress also enacted protections for providers who took action "in good faith" to restrict "obscene, lewd, lascivious, filthy, excessively violent, harassing or otherwise objectionable" content.[36] Under Section 230, information service providers are immune from liability both for hosting content that they know to be unlawful, and for removing content that they know to be constitutionally protected.[37]

Few other intermediary protections contain language quite as broad as Section 230. Compared with Section 230's broad immunity, "safe harbors" are a more common—and perhaps more justified—statutory approach, offering a conditional defense against liability.[38] Under Section 512 of the Digital Millennium Copyright Act of 1998, hosting providers are generally not liable for instances of copyright infringement by users so long as they do not know of the infringing material or activity.[39] In the European Union (EU), the E-Commerce Directive similarly shields service providers from liability for hosting users' illegal content so long as the providers do not have knowledge, authority, or control over the content.[40] Under both of these "safe harbor" provisions, online service providers are required to implement "notice-and-takedown" procedures to "expeditiously remove or disable access to" content alleged to be illegal.[41]

The decision to insulate platforms from liability for hosting user-generated speech reflects several political inclinations. First, it communicates

[34] *See* Klonick, *supra* note 10, at 1602.

[35] 47 U.S.C. § 230(c)(1) (2012).

[36] *Id.* § 230(c)(2).

[37] *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 333 (4th Cir. 1997); Hassell v. Bird, 420 P.3d 776, 793 (Cal. 2018) (holding that Section 230 immunity shielded Yelp from a court order directing it to take down defamatory consumer reviews); Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 408 (2017) ("Platforms have been protected from liability even though they republished content knowing it might violate the law, encouraged users to post illegal content, changed their design and policies for the purpose of enabling illegal activity, or sold dangerous products."). The breadth of Section 230's protections is not unlimited, however. Service providers have no immunity for violations of intellectual property law, federal criminal law, sex trafficking law, or the Electronic Communications Privacy Act of 1986, 47 U.S.C. § 230(e) (2012).

[38] James Grimmelmann, *Speech Engines*, 98 MINN. L. REV. 868, 946 n.371 (2014) (arguing that the safe harbor of Section 512 of the Digital Millennium Copyright Act of 1998 is "a better model" for search engine liability than Section 230 of the CDA under certain circumstances).

[39] 17 U.S.C. § 512(d)(1) (1998).

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

[41] 17 U.S.C. § 512(c)(1). *See also* E-Commerce Directive 2000/31, *supra* note 40, at art. 14 (containing nearly identical language).

a set of normative assumptions about the valuable role of private ordering in governing online speech.[42] Second, it relies upon a logic of stimulating innovation in order to expand the areas within which platforms were free to operate without government oversight.[43] Third, although this decision functionally created the space for platforms to self-govern, the primarily libertarian orientation towards Internet regulation—particularly in the U.S.—meant that the potential effects of corporate power and dominance were largely overlooked in favor of a focus on government censorship and surveillance.[44]

Likewise, many early advocates of Internet freedom recognized—and celebrated—the politics of online self-regulation, emphasizing the way in which the Internet would afford new autonomy to speakers and listeners without the existing constraints of government censorship.[45] The formation of online communities had radical democratic roots. Thinkers such as John Perry Barlow emphasized how the Internet infrastructure could invert the political economy of the telecommunications, media, and entertainment industries, describing the relationship between the production and consumption of information as being "as asymmetrical as that of bomber to bombee [sic]."[46] By creating spaces in which users created, curated, edited, and responded to content, the Internet could wrest control of that economy away from the media and entertainment industries.[47]

This vision of Internet freedom was often described as anarchic or "cyber libertarian."[48] It is perhaps more accurately described as "popu-

[42] *See* Bloch-Wehba, *supra* note 21, at 38 (describing how intermediary protections reflect a neoliberal approach to regulation).

[43] *See* Tarleton Gillespie, *The Politics of 'Platforms'*, 12 New Media & Soc'y 347, 356 (2010) (describing platforms' interest in "fostering a regulatory paradigm that gives them the most leeway to conduct their business . . . ."); Klonick, *supra* note 10, at 1607-08.

[44] *See* Zuboff, *supra* note 13, at 104 (describing "a few consistent themes: that technology companies such as Google move faster than the state's ability to understand or follow, that any attempts to intervene or constrain are therefore fated to be ill-conceived and stupid, that regulation is always a negative force that impedes innovation and progress, and that lawlessness is the necessary context for 'technological innovation.'"). *Cf.* Elizabeth Anderson, Private Government 40 (Princeton Univ. Press 2017) ("Should we not subject these forms of government to at least as much critical scrutiny as we pay to the democratic state?"); Zephyr Teachout & Lina Khan, *Market Structure and Political Law: A Taxonomy of Power*, 9 Duke J. Const. L. & Pub. Pol'y 37, 37 (2014).

[45] John Perry Barlow, *A Declaration of the Independence of Cyberspace*, EFF (Feb. 8, 1996), https://www.eff.org/cyberspace-independence [https://perma.cc/W5YL-4QDP] [hereinafter *A Declaration of the Independence of Cyberspace*].

[46] John Perry Barlow, *Death from Above*, Comm. ACM, May 1995, at 17, 17.

[47] *See* John Perry Barlow, *Property and Speech: Who Owns What You Say in Cyberspace?*, Comm. ACM, Dec. 1995, at 19, 20 (criticizing, in bombastic terms, the 1995 "White Paper," for extending copyright protection and contracting fair use, to the benefit of "media megacorps"). *See also* Michael Hauben & Ronda Hauben, *The Social Forces Behind the Development of Usenet*, First Monday, July 1998, https://firstmonday.org/ojs/index.php/fm/article/view/609/530 [https://perma.cc/C6MA-2Z77] ("The audience has very little choice over what is emphasized by most mass media. Usenet, however, is controlled by its audience.").

[48] Jack L. Goldsmith, *Against Cyberanarchy*, 65 U. Chi. L. Rev. 1199, 1203-04 (1998).

list," however, in the sense that it donned the mantle of popular support and vigorous opposition to elite interests.[49] Indeed, despite what Barlow called the "natural anarchy" of the Internet, he also acknowledged its distinctive forms of social order.[50] What he called the "unwritten codes" of online participation—the largely informal norms, rules, and policies that governed online services—were, in this telling, the expression of democratic self-governance, not instruments of censorship.[51]

Today, major platforms like Google, Facebook, and Twitter make rules that indelibly affect what, how, and where users are able to speak.[52] Platforms' moderation rules affect public discourse; information flow; and individual, free expression rights. Legal scholars who analyze these issues from the perspective of free expression often see the substantive rules of content moderation as performing an important, law-like function, setting the boundaries of participation in an online community and the penalties for non-compliance with those rules.[53]

That intermediary immunities that have created breathing room for platforms to create their own, voluntary, quasi-regulatory speech constraints may seem ironic in light of the freewheeling libertarianism of early Internet freedom advocates. But the proliferation of monitoring, filtering, and moderation technologies is the direct result of intermediary immunities and safe harbors created to stimulate innovation.[54] As Annemarie Bridy pointed out, Section 230's model of intermediary immunity both allows platforms to take down speech that public actors could not censor, and "frees them to develop and experiment with new tools for doing so, including automated technical measures."[55]

[49] *See, e.g.*, Margaret Canovan, *Trust the People! Populism and the Two Faces of Democracy*, 47 Pol. Stud. 2, 4 (1999) ("Populists claim legitimacy on the grounds that they speak for *the people*: that is to say, they claim to represent the democratic sovereign, not a sectional interest such as an economic class."); Cas Mudde, *The Populist Zeitgeist*, 39 Gov't & Opposition 541, 541, 544 (2004).

[50] *See* John Perry Barlow, *The Great Work*, Comm. ACM, Jan. 1992, at 25, 25–26.

[51] *A Declaration of the Independence of Cyberspace*, *supra* note 45.

[52] *See* Marvin Ammori, *The "New"* New York Times*: Free Speech Lawyering in the Age of Google and Twitter*, 127 Harv. L. Rev. 2259, 2260 (2014); Chander, *supra* note 10, at 1809, 1816; Klonick, *supra* note 10, at 1616–17; Sarah Myers West, *Censored, Suspended, Shadowbanned: User Interpretations of Content Moderation on Social Media Platforms*, 20 New Media & Soc'y 4366, 4366 (2018) [hereinafter *Censored, Suspended, Shadowbanned*].

[53] Klonick, *supra* note 10, at 1630; Rory Van Loo, *The Corporation as Courthouse*, 33 Yale J. on Reg. 547, 602 (2016).

[54] Kate Klonick, *Why the History of Content Moderation Matters*, Techdirt (Jan. 30, 2018, 11:55 AM), https://www.techdirt.com/articles/20180129/21074939116/why-history-content-moderation-matters.shtml [https://perma.cc/E7G2-UQ4V] ("[M]ore important than understanding the intricacies of the system is understanding the history of how it was developed.").

[55] Annemarie Bridy, *Leveraging CDA 230 to Counter Online Extremism*, Geo. Wash. Program on Extremism (Sept. 2019), https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/Leveraging%20230%20to%20Counter%20Online%20Extremism.pdf [https://perma.cc/8SKA-KHNW].

B. Spam Filtering

Made immune from legal responsibility for user-generated content, web platforms could have become a free-for-all.[56] But the "superabundant resources" of the web were not inexhaustible.[57] The burgeoning number of web users put stress on some of the early web services.[58] Trolling,[59] spam,[60] manipulation,[61] and other kinds of misbehavior grew common. These kinds of misbehavior had the potential to damage online communities and drained resources. Online communities reacted by developing their own rules and restrictions to both, constrain online misbehavior, and promote desirable behavior.

At least in theory, early Internet platforms lent themselves to forms of moderation from the bottom up. For instance, Usenet newsgroups,—an archetypal example of the dynamics of early cyberspace—were decentralized and could either be moderated or unmoderated.[62] But whether moderated or not, Usenet was celebrated for its "uncensored" nature precisely because it was cooperative and "controlled by its audience" rather than a central authority.[63] For Usenet, the power to moderate content went hand in hand with the power to constitute a "community of interest." In other words, the power to engage in the explicitly political act of defining the limits of acceptable behavior in a social group.[64]

Usenet's experience with spam shows exactly how, despite its decentralization, the grassroots approach to online moderation in fact embraced order over chaos. In 1994, two lawyers, Laurence Canter and Marsha

[56] Klonick, *supra* note 10, at 1604.

[57] Milton L. Mueller, Networks and States: The Global Politics of Internet Governance 2 (MIT Press 3d ed. 2010).

[58] Ed Krol, *It's Time to Give Usenet a Much-Needed Overhaul*, Network World, Apr. 1, 1996, at 65 (describing how, as Usenet grew, discussion groups got "too active" to be sustainable).

[59] Mattathias Schwartz, *The Trolls Among Us*, N.Y. Times (Aug. 3, 2008), https://www.nytimes.com/2008/08/03/magazine/03trolls-t.html [https://perma.cc/2M4C-GP2R].

[60] Finn Brunton, Spam: A Shadow History of the Internet xvi (MIT Press 2013) (describing how spammers work "often by directly exploiting the same technologies and beneficial effects that enable the communities on which they predate").

[61] Sara Kiesler et al., *Regulating Behavior in Online Communities*, *in* Building Successful Online Communities: Evidence-Based Social Design 125, 128 (Robert E. Kraut & Paul Resnick eds., MIT Press 2011).

[62] *See, e.g.*, Jeffrey M. Taylor, *Liability of Usenet Moderators for Defamation Published by Others: Flinging the Law of Defamation into Cyberspace*, 47 Fla. L. Rev. 247, 254 (1995). *See also* Martin Dodge & Rob Kitchin, Mapping Cyberspace 135 (Routledge 1st ed. 2001) ("Usenet can be thought of as the archetype of uncontrolled cyberspace, since it is highly distributed, free-wheeling, and has no official funding, external quality control or censorship.").

[63] *See* Hauben & Hauben, *supra* note 47. *See also* Janet Abbate, Inventing the Internet 203 (MIT Press 1999) (describing how commercial providers began to "imitate" grassroots systems like Usenet through proprietary protocols).

[64] *See* Abbate, *supra* note 63, at 201. *See also* Anne Wells Branscomb, *Anonymity, Autonomy, and Accountability: Challenges to the First Amendment in Cyberspaces*, 104 Yale L.J. 1639, 1656–58 (1995) (describing how "netizens" have chosen to use moderation through voting).

Siegel, sent mass messages to Usenet newsgroups with the subject line "Green Card Lottery - Final One?" touting their own legal services—a massive violation of Usenet norms against unsolicited email.[65] The reaction was hostile, to say the least. Recipients sent "'electronic letter bombs' designed to destroy" the ads,[66] and created an electronic beeper which called the Canter & Siegel law offices repeatedly during the night, filling their voicemail boxes.[67]

The Canter & Siegel incident became an infamous turning point for online communications. Canter has been called "the father of modern spam,"[68] and the Green Card Lottery spam message is widely described as the "public debut" of unsolicited commercial advertising on Usenet.[69] But rather than being the death of grassroots moderation, the Canter & Siegel advertisement showed the flexibility and adaptability of the Usenet community. In part, Usenet users responded to the Canter & Siegel advertisement by organizing the community to implement *ex ante* screening protocols to identify and flag, or delete, suspected spam.[70]

Canter and Siegel also showed the power and breadth of online commercial advertising at the very moment that commercialization and privatization of the Internet began to pick up speed.[71] While the Internet's backbone had been formally transferred from military to civilian control in 1990, the National Science Foundation continued to run the backbone, which specifically prohibited commercial activities.[72] Throughout the early and mid-1990s, however, new commercial network service providers emerged, creating a new commercial infrastructure for the Internet that ultimately replaced the old government-run backbone.[73] In a 2002 interview, Canter reported that the Green Card Lottery incident had caused several service providers to terminate the Canter & Siegel account because their servers lacked the capacity for the "huge amounts of traffic" that the advertisement generated.[74]

As commercial infrastructure improved, and commercial email service providers emerged, spam remained a universal annoyance. As a result, the

[65] Branscomb, *supra* note 64, at 1656–58; Lorrie Faith Cranor & Brian A. LaMacchia, *Spam!*, Comm. ACM, Aug. 1998, at 74, 75.

[66] *See* Branscomb, *supra* note 64, at 1658.

[67] *Id.* at 1658 n.70.

[68] *The Father of Modern Spam Speaks*, CNET (Mar. 26, 2002, 12:19 PM), https://www.cnet.com/news/the-father-of-modern-spam-speaks/ [https://perma.cc/WM7Z-QZJV].

[69] Cranor & LaMacchia, *supra* note 65, at 74. *See also* Brian Hayes, *Computing Science: Spam, Spam, Spam, Lovely Spam*, 91 Am. Scientist 200, 200–01 (2003).

[70] *See* Hayes, *supra* note 69, at 200–01. *See also* Brunton, *supra* note 60, at 96 (describing the formation of the news.admin.net-abuse.email newsgroup, or NANAE, to respond to spam).

[71] *See The Father of Modern Spam Speaks*, *supra* note 68 ("What we definitely showed was that you could reach a lot of people—huge numbers of people! Today it would be the equivalent to reaching millions relatively easily.").

[72] Abbate, *supra* note 63, at 196 (explaining that "Congress was quick to condemn any use of government-subsidized resources for commercial purposes").

[73] *See id.* at 197–99.

[74] *The Father of Modern Spam Speaks*, *supra* note 68.

technology of spam filtering developed rapidly.[75] In August 2002, Paul Graham published *A Plan for Spam*, an essay that endorsed the use of naïve, Bayesian, statistical analysis for spam filtering.[76] Graham's approach relied on a statistical analysis of tokens in two corpus—one of spam, and one of non-spam—to determine the probability that a given message was spam.[77] Graham's statistical analysis laid the groundwork for spam filtering technology that could adapt quickly over time as spammers deployed new language to circumvent filters.[78]

Although governments also acted, the private sector proved more effective at enforcing anti-spam measures. When Congress enacted the CAN-SPAM Act in 2003, it regulated the structure of spam messages and certain methods used to send them.[79] At the end of the day, though, CAN-SPAM proved difficult to enforce.[80] Code-based spam filters, however, have dramatically improved the experience of email users.[81] Technology companies have invested heavily in filtering technologies for spam, which accounted for over 90% of all email by 2009.[82] Platforms have also extended spam filtering techniques to other categories of bad content online.[83]

1. *"Artificial Intelligence"*

Usenet's response to Canter & Siegel demonstrated the power of social norms as a moderating influence.[84] But as platforms grew, expanded, and commercialized, social norms provided less cohesion.[85] Graham's innovation illustrated how the social and legal norms of online content modera-

[75] Brunton, *supra* note 60, at 126-28 (describing spam research).

[76] *A Plan for Spam*, Paul Graham (Aug. 2002), http://www.paulgraham.com/spam.html [https://perma.cc/6VB8-4PSP] (last visited Mar. 2, 2020) [hereinafter Graham]. *See also* Brunton, *supra* note 60, at 133-35.

[77] Graham, *supra* note 76.

[78] Brunton, *supra* note 60, at 140-41.

[79] Roger Allan Ford, *Preemption of State Spam Laws by the Federal CAN-SPAM Act*, 72 U. Chi. L. Rev. 355, 358-60 (2005).

[80] *Id.* at 356.

[81] Bradley Taylor et al., *The War Against Spam: A Report from the Front Line*, *in* NIPS 2007 Workshop on Machine Learning Adversarial Environments for Computer Security 1, 1, 3 (2007).

[82] *See generally* Sarita Yardi et al., *Detecting Spam in a Twitter Network*, First Monday, Jan. 2010, https://firstmonday.org/ojs/index.php/fm/article/view/2793 [https://perma.cc/K4FY-BPZY]. *See also Spam (Unsolicited Commercial E-Mail): Hearing Before the Comm. on Commerce, Sci., & Transp.*, 108th Cong. 3 (2003) (opening statement of Hon. John McCain, Chairman of the Committee).

[83] *See infra* Section III.E.

[84] Justin Peters, *Original Sin: The Creation of Email Spam and Its Threat to the Promise of the Internet*, Colum. Journalism Rev. (2013), https://archives.cjr.org/critical_eye/original_sin.php [https://perma.cc/5EKF-HSKZ].

[85] Caitlin McLaughlin & Jessica Vitak, *Norm Evolution and Violation on Facebook*, 14 New Media & Soc'y 299, 300 (2012) ("Because of the speed at which these sites have evolved, however, an established set of social norms guiding users' behavior has been slow to follow. Furthermore, when behavioral norms are ambiguous, it becomes more difficult to both establish a formal set of norms and to respond to perceived norm violations.").

tion could be encoded into architecture and design. The emerging availability of filtering protocols and software meant that Usenet-style social norms against spam could be formalized into code deployed at the platform or provider level without draining platform resources.

As commercialized platforms developed an increasingly robust set of rules and policies to guide participation on the platforms, they embraced automated filtering as a critical tool for scaling the application of these standards. This focus on scale is partly responsible for the transformation that moderation has undergone from the grassroots of the early Internet into the more "industrial" version we see today: moderation techniques that rely heavily on both, *ex ante*, automated screening mechanisms, as well as *ex post* review by human moderators (equally aided by machines).[86]

Platforms seized the opportunity to use automated, *ex ante* screening to exclude spam from their services, but their different business models also meant that definitions of spam, and automated techniques to address it, diverged.[87] Gmail's program policies prohibit—but do not define—"spam," and remind users to "keep in mind that [their] definition of 'unsolicited' or 'unwanted' mail may differ from [the] email recipients' perception."[88] In contrast, Facebook's community standards prohibit "commercial spam," and explicitly instruct users not to "artificially increase distribution for financial gain."[89] YouTube's community guidelines prohibit "spam, scams, and other deceptive practices that take advantage of the YouTube community," including voter suppression.[90] These differences not only demonstrate that the very definition of prohibited "spam" can vary greatly even between major mainstream communities, but also that the technical and quasi-legal architecture of filtering can flexibly accommodate different kinds of restrictions.

Intermediary immunities and safe harbor protections allowed platforms to disclose their content-filtering mechanisms. Even so, companies have struggled to ensure that their abuse detection "scales" to meet the needs of global communication.[91] The problem is that "the more ambiguous and contextual classificatory criteria become, the more difficult it

[86] Robyn Caplan, *Content or Context Moderation?: Artisinal, Community-Reliant, and Industrial Approaches*, DATA & SOC'Y (Nov. 14, 2018), https://datasociety.net/library/content-or-context-moderation/ [https://perma.cc/V74Z-P4AM]. *See also* Gorwa et al., *supra* note 17, at 7–10.

[87] *Cf. The Virtues of Moderation*, *supra* note 28, at 67–68 (distinguishing between *ex ante* and *ex post*).

[88] *Gmail Program Policies*, GOOGLE, https://www.google.com/gmail/about/policy/ [https://perma.cc/D8NH-X8AD] (last visited Feb. 5, 2020).

[89] *Community Standards: Spam*, FACEBOOK, https://www.facebook.com/communitystandards/spam [https://perma.cc/E2KM-2QQC] (last visited Feb. 5, 2020).

[90] *Spam, Deceptive Practices & Scams Policies*, YOUTUBE, https://support.google.com/youtube/answer/2801973?hl=en [https://perma.cc/L4YJ-4R6H] (last visited Feb. 5, 2020).

[91] Taylor et al., *supra* note 81, at 3 (describing the challenges of internationalizing machine learning for spam filters).

becomes to train algorithms accurately."[92] Compared to contextual decisions about whether a depiction of violence is a piece of extremist content or a human rights report about extremism, spam is relatively easy. For example, Bayesian filters would respond as spammers started to use "v14gr4" instead of "Viagra."[93]

While platforms often tout the sophistication of their machine learning methods, these methods often falter in significant ways. Consider Perspective API (Perspective), a popular machine learning system developed by Google and Jigsaw to help combat trolling and "improve conversations online."[94] In 2017, University of Washington researchers demonstrated how easily Perspective could be fooled, finding that "an adversary can subtly modify a toxic phrase"—for example, by misspelling the word "idiots" as "idiiots," "id.iots," or "i.diots"—to significantly lower the "toxicity score" and the likelihood that a comment would be classified as "rude" or trolling.[95] This year, researchers also demonstrated that Perspective disproportionately identifies posts written in African-American Vernacular English as "rude" or "toxic," reflecting—and amplifying—racial bias.[96]

Despite their drawbacks, machine learning and artificial intelligence are critical tools for scaling content moderation. Indeed, stemming the tide of bad content became a matter of corporate survival. Spam filtering illustrated the potential benefits automation held for both, users and platforms seeking to limit certain kinds of online content. Today, platforms employ a variety of techniques to make content-related decisions far beyond spam. Just as Graham's *Plan for Spam* approach prescribed Bayesian analysis to help an algorithm predict the likelihood that a given email was spam, modern automated techniques often use machine learning algorithms to predict the likelihood that a piece of content violates the platforms' rules or the law.[97] Like spam, much of this content, while undesirable, is not *ille-*

[92] Kirsten Gollatz et al., *The Turn to Artificial Intelligence in Governing Communication Online*, HIIG 1, 7 (2018), https://www.hiig.de/wp-content/uploads/2018/09/Workshop-Report-2018-Turn-to-AI.pdf [https://perma.cc/VUS2-HLAP].

[93] GRAHAM, *supra* note 76 (explaining his optimism that Bayesian filters would respond as spammers started to use "'c0ck' instead of 'cock'").

[94] *See generally* PERSPECTIVE, https://www.perspectiveapi.com/#/home [https://perma.cc/2GAY-5X77] (last visited Jan. 31, 2020).

[95] Hossein Hosseini et al., *Deceiving Google's Perspective API Built for Detecting Toxic Comments*, ARXIV 1, 2 tbl.1 (2017), https://arxiv.org/pdf/1702.08138.pdf [https://perma.cc/TV3A-9CBG].

[96] Maarten Sap et al., *The Risk of Racial Bias in Hate Speech Detection*, *in* PROCEEDINGS OF THE 57TH ANNUAL MEETING OF THE ASSOCIATION OF COMPUTATIONAL LINGUISTICS 1668, 1668–70, 1677 (ACL 2019); Anna Chung, *How Automated Tools Discriminate Against Black Language*, MEDIUM (Feb. 28, 2019), https://onezero.medium.com/how-automated-tools-discriminate-against-black-language-2ac8eab8d6db [https://perma.cc/5EZG-TVWP]. *See also* Thomas Davidson et al., *Racial Bias in Hate Speech and Abusive Language Detection Datasets*, ARXIV 1, 5–8 (2019), https://arxiv.org/pdf/1905.12516v1.pdf [https://perma.cc/V3YG-R7CZ] (demonstrating "substantial racial disparities" in the performance of language classifiers intended to detect hate speech and abusive language).

[97] *See* Natasha Duarte et al., *Mixed Messages? The Limits of Automated Social Media Content Analysis*, CTR. DEMOCRACY & TECH. (Nov. 28, 2017), https://cdt.org/insights/mixed-messages-the-limits-of-automated-social-media-content-analysis/ [https://

gal. Platforms use keyword filters to exclude hashtags that promote disordered eating;[98] analyze the content of photographs to determine whether they include "adult content;"[99] and exclude posts that use politically sensitive terms.[100]

C. Unlawful Content

Automated, proactive screening is not limited to legal content that platforms simply find distasteful, too resource-intensive, or would otherwise prefer not to host. Technology companies have also developed tools that use fingerprinting and hashing technologies to flag and limit the distribution of illegal content. Many of the content-related decisions that platforms seek to automate often require more context and judgment than determining whether an email should go to the recipient's inbox or junk folder, and also have higher stakes.

Unlike Bayesian filtering or other content-based automated screening techniques, which predict the likelihood that a piece of content should be taken down, fingerprinting and hashing technologies essentially work by screening the characteristics of user-uploaded content against an existing database of characteristics that indicate illegality.[101] In order to screen user-generated content, fingerprinting and hashing technologies require a library of content that has *already* been determined to possess the relevant characteristics.[102]

Many platforms rely on sophisticated hashing technology to identify and prevent the re-upload of specific child sexual abuse images.[103] "Hashing" means to apply a mathematical function that generates a series of characters to identify a given input.[104] For example, one might use a hash function to generate a string of characters to identify a photograph, a text

perma.cc/M8FS-R2XM] (outlining how spam filtering gave rise to different natural language processing techniques).

[98] Stevie Chancellor et al., *#Thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities*, *in* Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing 1201 (ACM 2016); Ysabel Gerrard, *Beyond the Hashtag: Circumventing Content Moderation on Social Media*, 20 New Media & Soc'y 4492, 4494–96 (2018).

[99] Shannon Liao, *Tumblr Will Ban All Adult Content on December 17th*, Verge (Dec. 3, 2018, 12:26 PM), https://www.theverge.com/2018/12/3/18123752/tumblr-adult-content-porn-ban-date-explicit-changes-why-safe-mode [https://perma.cc/EKL5-CHUG].

[100] *See generally* David Bamman et al., *Censorship and Deletion Practices in Chinese Social Media*, First Monday, Mar. 2012, https://firstmonday.org/ojs/index.php/fm/article/download/3943/3169 [https://perma.cc/S2TH-773X].

[101] *See generally* Gorwa et al., *supra* note 17.

[102] *Id.* at 2.

[103] *See* Jeff Kosseff, *Private Computer Searches and the Fourth Amendment*, 14 I/S: J.L. & Pol'y for Info. Soc'y 187, 209 (2018). *See, e.g.*, *PhotoDNA*, Microsoft, https://www.microsoft.com/en-us/photodna [https://perma.cc/J5MV-Y7FU] (last visited June 21, 2020).

[104] Dennis Martin, Note, *Demystifying Hash Searches*, 70 Stan. L. Rev. 691, 695 (2018).

file, or the contents of a hard drive.[105]

PhotoDNA, a tool developed by Microsoft and licensed for free to technology companies and law enforcement, can match the hash values of photos or videos uploaded by individual users against a database of hash values of other photos or videos containing illegal images of child sexual abuse.[106] If PhotoDNA finds a match between user-generated content and known child sexual abuse imagery, the software sends a "CyberTip" directly to the National Center for Missing and Exploited Children.[107] Using PhotoDNA, law enforcement can "identify child pornography with almost absolute certainty, regardless of the name associated with a file."[108]

Technology companies are not required by law to use proactive monitoring or filtering to detect child sexual abuse imagery, but rather have done so voluntarily.[109] Under the Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today Act of 2003 (PROTECT Act), companies that obtain "actual knowledge" of child pornography are required to report it to the National Center for Missing and Exploited Children.[110] Despite being shielded from liability, platforms have undertaken extensive voluntary action to limit illegal online content, as the PhotoDNA example illustrates. An online service provider might develop these programs for its own purposes, such as to "protect its own business and reputation and to protect the users" of its systems.[111]

Similarly, platforms are also adopting methods of proactive, *ex ante* screening for violent extremist and terrorist content in response to national and global pressures.[112] In 2016, in response to pressure by European governments, several major technology companies formed a consortium, the Global Internet Forum to Counter Terrorism (GIFCT), and announced that they would create a "shared industry database of 'hashes'—unique digital 'fingerprints'—for violent terrorist imagery or terrorist recruitment videos or images."[113] The hash database deploys a technology similar to that used in countering child sexual abuse imagery: participants in the effort build a database of hash values that serve as identifiers for files

[105] Richard P. Salgado, *Fourth Amendment Search and the Power of the Hash*, 119 Harv. L. Rev. F. 38, 39 (2006) (replying to Orin S. Kerr, *Searches and Seizures in a Digital World*, 119 Harv. L. Rev. 531 (2005)).

[106] *See generally id.*; *PhotoDNA*, *supra* note 103.

[107] United States v. Reddick, 900 F.3d 636, 637–38 (5th Cir. 2018).

[108] United States v. Larman, 547 F. App'x 475, 477 (5th Cir. 2013).

[109] 18 U.S.C. § 2258A(a)(1) (2008); Susan Klein & Crystal Flinn, *Social Media Compliance Programs and the War Against Terrorism*, 8 Harv. Nat'l Sec. J. 53, 78–79 (2017).

[110] Klein & Flinn, *supra* note 109, at 78 (citing to PL 98-473, 98 Stat. 1837 (1984), codified at 42 U.S.C. § 5773(b) (2015)).

[111] United States v. Green, 857 F. Supp. 2d 1015, 1018 (S.D. Cal. 2012) (quoting testimony of Don Colcolough, AOL's Director of Investigations and Cyber Security). *See also* United States v. Keith, 980 F. Supp. 2d 33, 40 (D. Mass. 2013) (finding that AOL had "an important business reason" for its Image Detection and Filtering Process, or IDFP).

[112] *See generally* Bloch-Wehba, *supra* note 21.

[113] *Partnering to Help Curb the Spread of Terrorist Content Online*, Google (Dec. 5, 2016), https://blog.google/topics/google-europe/partnering-help-curb-spread-terrorist-content-online/ [https://perma.cc/5Z6H-P4Z3].

known to correspond to violent extremist or terrorist content.[114]

Though platforms are not required to screen for "terrorist" content, they have proudly advertised their abilities to do so. For instance, Facebook expressed support for automated content deletion for terrorist content before the European Commission, noting that the platform had removed 99% of ISIS and Al-Qaeda terror content before it had been flagged by users.[115] YouTube's most recent report, documenting the enforcement of its Community Guidelines, likewise states that "automated flagging enables us to act more quickly and accurately to enforce our policies."[116]

Platforms' proactive initiatives are largely a response to escalating threats of regulatory action by the European Commission.[117] In the EU, the Terrorism Directive explicitly calls for exploring the possibility of "voluntary action" by platforms or by state actors to "detect[ ] and flag[ ]" terrorist content online pursuant to platforms' terms of service.[118] The European Commission was dissatisfied with platforms' approach to proactive filtering and followed up with a recommendation on "measures to effectively tackle illegal content online."[119] The recommendation exhorted platforms to take "proportionate and specific proactive measures, including by using automated means," to find, remove, and prevent the reposting of terrorist content.[120] Recognizing that the use of automated filtering would be difficult for smaller platforms, the Commission also "encourage[d]" platforms to "cooperate" in sharing technological tools to curb terrorist content.[121]

Until spring of 2019, political pressures to address violent extremism were largely limited to Islamic terrorism. That changed, however, after the March 2019 shootings at the Masjid al Noor and Linwood Islamic Centre in Christchurch, New Zealand prompted a wave of responses from platforms and government actors reconsidering the role of social media in

[114] *See* Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money* 6 (Hoover Inst. Aegis Paper Series No. 1807, 2018) [hereinafter *Internet Platforms*]; T.J. McIntyre, *Child Abuse Images and Cleanfeeds: Assessing Internet Blocking Systems*, *in* Research Handbook on Governance of the Internet 277, 288 (Ian Brown ed., Edward Elgar 2013).

[115] European Commission Press Release IP/17/5105, Fighting Terrorism Online: Internet Forum Pushes for Automatic Detection of Terrorist Propaganda (Dec. 6, 2017) (Monika Bickert, Director of Global Policy Management, Facebook, stated that "[t]he use of AI and other automation to stop the spread of terrorist content is showing promise.").

[116] *YouTube Community Guidelines Enforcement*, Google, https://transparencyreport.google.com/youtube-policy/removals?hl=en [https://perma.cc/2YWV-5SSH] (last visited Mar. 2, 2020).

[117] Bloch-Wehba, *supra* note 21, at 43.

[118] Directive 2017/541, of the European Parliament and of the Council of 15 March 2017 on Combating Terrorism and Replacing Council Framework Decision 2002/475/JHA and Amending Council Decision 2005/671/JHA, 2017 O.J. (L 88) 6, 9 (EU).

[119] *Commission Recommendation of 1 March 2018 on Measures to Effectively Tackle Illegal Content Online*, at 1, 2, C(2018) 1177 final (Mar. 1, 2018) (EC).

[120] *Id.* at 8.

[121] *Id.*

amplifying hatred and violence.[122] The shooter, who streamed on Facebook Live for seventeen minutes as he killed fifty-one people, had released a white supremacist manifesto that continued to circulate online well after New Zealand banned its possession.[123] Facebook, YouTube, and Instagram faced questions about why and how restricted footage and imagery from the shooting continued to resurface on their platforms.[124]

While Facebook had historically banned white supremacist content, it did not take steps to eliminate white nationalism and white separatism from the platform until after Christchurch.[125] YouTube similarly changed its policy in June 2019 to prohibit "videos alleging that a group is superior in order to justify discrimination, segregation or exclusion based on qualities like age, gender, race, caste, religion, sexual orientation or veteran status."[126] Christchurch came in the midst of a string of other mass killings linked to white nationalism. In the U.S., mass murderers at the Tree of Life Synagogue in Pittsburgh, Pennsylvania; the Chabad of Poway, California; and the Walmart in El Paso, Texas all posted white nationalist manifestos on social media. The Poway and El Paso shooters explicitly cited Christchurch as inspiration.[127]

Governments and platforms appeared to re-double their efforts to address violent extremism after Christchurch. In May 2019, New Zealand and France led a meeting of government actors, technology companies, and civil society organizations at which they adopted the Christchurch Call to Action to Eliminate Terrorist and Violent Extremist Content Online (Christchurch Call). The Christchurch Call is a non-binding agreement committing the signatories—dozens of nations and major technology companies including Amazon, Facebook, Google, Microsoft, Twitter, and YouTube—to "work collectively" to "counter violent extremism in all its forms" and to "accelerate research into and development of technical solutions to prevent the upload of and to detect and immediately remove terrorist and

[122] *See Inside the Team at Facebook That Dealt with the Christchurch Shooting*, *supra* note 3 (describing how content moderation teams struggled to keep up with the footage).

[123] Charles Anderson, *Censor Bans 'Manifesto' of Christchurch Mosque Shooter*, GUARDIAN (Mar. 23, 2019, 10:53 PM), https://www.theguardian.com/world/2019/mar/24/censor-bans-manifesto-of-christchurch-mosque-shooter [https://perma.cc/94UM-4TLG]; Charlotte Graham-McLay, *Spreading the Mosque Shooting Video Is a Crime in New Zealand*, N.Y. TIMES (Mar. 21, 2019), https://www.nytimes.com/2019/03/21/world/asia/new-zealand-attacks-social-media.html [https://perma.cc/TLP2-8G3S].

[124] James Rogers, *Horrific Footage of Christchurch Mosque Shooting Surfaces on YouTube and Instagram*, FOX NEWS (Aug. 6, 2019), https://www.foxnews.com/tech/footage-christchurch-mosque-shooting-youtube-instagram [https://perma.cc/994Z-JZMP].

[125] *Standing Against Hate*, FACEBOOK (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate/ [https://perma.cc/Y24X-NQJ4].

[126] *Our Ongoing Work to Tackle Hate*, YOUTUBE (June 5, 2019), https://youtube.googleblog.com/2019/06/our-ongoing-work-to-tackle-hate.html [https://perma.cc/6WZV-Y5GP].

[127] Tim Arango et al., *Minutes Before El Paso Killing, Hate-Filled Manifesto Appears Online*, N.Y. TIMES (Aug. 3, 2019), https://www.nytimes.com/2019/08/03/us/patrick-crusius-el-paso-shooter-manifesto.html [https://perma.cc/K4TK-B5J9].

violent extremist content online."[128]

At the same time, GIFCT publicly committed to a range of action steps to implement the Christchurch Call goals, including investing in AI-based and fingerprinting technologies to "detect and remove terrorist and violent extremist content."[129] A few months later, GIFCT announced that it would become its own independent organization, a step that would allow it to "do even more"—but which prompted critics to wonder about the continuing influence of dominant platforms.[130]

While the threat of regulation is a powerful motivation for platforms to synchronize their actions with law enforcement goals, platforms also have significant business incentives to voluntarily prevent their services from being used to spread horrific illegal content. And in the context of illegal, online content, automated flagging has proven to be a powerful opportunity for collaboration and cooperation between the private and public sectors. In fact, as with many forms of cybercrime, close cooperation between the government and private sector is critical for successful prosecution.[131]

These concerns have been particularly pronounced because of the emergence of "Internet Referral Units (IRUs)," which are specialized police units that monitor online activity for terms of service violations and cybercrime.[132] Several major companies have partnered with these units as "trusted flaggers" whose complaints receive expedited treatment.[133] IRUs' use of the mechanisms of private governance for law enforcement purposes has raised concerns about the transparency, accountability, and redress mechanisms for censorship.[134]

[128] *Christchurch Call to Eliminate Terrorist & Violent Extrimist Content Online*, CHRISTCHURCH CALL (2019), https://www.christchurchcall.com/christchurch-call.pdf [https://perma.cc/4DVS-RZPJ].

[129] *Actions to Address the Abuse of Technology to Spread Terrorist and Violent Extremist Content*, GIFCT (May 15, 2019), https://gifct.org/press/actions-address-abuse-technology-spread-terrorist-and-violent-extremist-content/ [https://perma.cc/TB7T-W874]; Ángel Díaz, *Global Internet Forum to Counter Terrorism's 'Transparency Report' Raises More Questions Than Answers*, JUST SEC. (Sept. 25, 2019), https://www.justsecurity.org/66298/gifct-transparency-report-raises-more-questions-than-answers/ [https://perma.cc/BH5S-EVYJ]; Andrew Sullivan, *Looking the GIFCT in the Mouth*, INTERNET SOC'Y (Oct. 11, 2019), https://www.internetsociety.org/blog/2019/10/looking-the-gifct-in-the-mouth/ [https://perma.cc/NT9B-UBAU]. *See also* Emma Llansó, *Platforms Want Centralized Censorship. That Should Scare You*, WIRED (Apr. 18, 2019, 9:00 AM), https://www.wired.com/story/platforms-centralized-censorship/ [https://perma.cc/GN69-GZYL].

[130] *Next Steps for GIFCT*, GIFCT (Sept. 23, 2019), https://gifct.org/press/next-steps-gifct/ [https://perma.cc/X2YR-WUZJ].

[131] *See, e.g.*, Kristen E. Eichensehr, *Public-Private Cybersecurity*, 95 TEX. L. REV. 467, 474 (2017).

[132] *See* Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114, 120–22 (2018).

[133] Bloch-Wehba, *supra* note 21, at 45–46.

[134] Jason Pielemeier & Chris Sheehy, *Understanding the Human Rights Risks Associated with Internet Referral Units*, MEDIUM (Feb. 25, 2019), https://medium.com/global-network-initiative-collection/understanding-the-human-rights-risks-associated-with-internet-referal-units-by-jason-pielemeier-b0b3feeb95c9 [https://perma.cc/A8CJ-AL5J].

Despite the increasingly intertwined relationship between business and law enforcement interests, platforms' powerful role is subject to little oversight or accountability. Many of the most powerful techniques to address crime, such as the terrorism and child sexual abuse hash databases, are formally private and voluntary.[135] These techniques are specifically designed to ferret out unlawful content, and in the case of child sexual abuse imagery, to actually report it to the police.[136] However, there is scant judicial or public oversight of these practices despite the close relation between platform reporting and law enforcement interests. Moreover, technology companies rarely seem to consider the role of their business models in whetting the public's appetite for the very unlawful or undesirable content that they seek to suppress.[137]

D. Copyright Enforcement

Ex ante content recognition technologies developed by the private sector have also transformed copyright enforcement. In response to copyright takedown requests, companies have responded with a range of technical approaches, including facilitating site-wide deletion of content,[138] and employing hashing or fingerprinting to identify and filter infringing content.[139] These technologies have fundamentally altered intellectual property enforcement online.[140]

Under Section 512 of the Digital Millennium Copyright Act of 1998 (DMCA), online service providers are generally immune from secondary liability for copyright infringement when users transmit infringing material through their platforms, unless the provider has "actual knowledge" or "awareness" of the infringing content.[141] Europe has historically also embraced safe harbors for intermediary service providers that host user-generated content. Under the European Commission's E-Commerce Direc-

[135] *See* discussion *supra* Section I.C.

[136] United States v. Rosenschein, No. CR 16-4571 JCH, 2019 WL 4855428, at *6 (D.N.M. Oct. 2, 2019). In the United States, numerous federal courts have held that online service providers are not government "agents." But in one case, discovery as to the nature of the relationship between PhotoDNA and law enforcement is still ongoing. *Id.* (agreeing to "honor Microsoft's commitment to provide agreements pertaining to PhotoDNA and any federal law enforcement agency or State Attorney General, and agreements with third parties regarding hash sharing."). *See also* United States v. Stevenson, 727 F.3d 826, 831 (8th Cir. 2013); United States v. Richardson, 607 F.3d 357, 367 (4th Cir. 2010).

[137] *See* Julia Alexander, *YouTube Still Can't Stop Child Predators in Its Comments*, VERGE (Feb. 19, 2019, 12:50 PM), https://www.theverge.com/2019/2/19/18229938/youtube-child-exploitation-recommendation-algorithm-predators [https://perma.cc/4T4P-2E4S].

[138] *Copyright Management Tools*, YOUTUBE, https://support.google.com/youtube/answer/9245819?hl=en&ref_topic=9282364 [https://perma.cc/2AYX-F6A4] (last visited Jan. 31, 2019).

[139] Alexander, *supra* note 137.

[140] Annemarie Bridy, *Graduated Response and the Turn to Private Ordering in Online Copyright Enforcement*, 89 OR. L. REV. 81, 93–94 (2010); Annemarie Bridy, *Internet Payment Blockades*, 67 FLA. L. REV. 1523, 1538 n. 100 (2015) (citing 17 U.S.C. § 512 (a)–(d) (2012)).

[141] 17 U.S.C. § 512(c)(1)(A) (1999).

tive, member states may not hold intermediary service providers liable for content posted by users, so long as providers lack "actual knowledge of illegal activity or information," or "act[ ] expeditiously to remove or to disable access" once they gain knowledge.[142]

Both the European and U.S. regimes are reactive, not proactive, and emphasize the need for timely deletion upon request: a "notice and takedown" regime.[143] Providers generally only "know" of infringement once a copyright holder notifies the provider that a specific piece of content infringes their copyright. Both Section 512 of the DMCA and the E-Commerce Directive explicitly disavow the intention to require service providers to monitor content that was hosted for or generated by users for illegality.[144] Under the DMCA, service providers have no statutory obligations to "monitor" their platforms or "affirmatively seek[ ] facts indicating infringing activity," unless doing so is a "standard technical measure."[145] The E-Commerce Directive likewise bars member states from imposing general obligations on intermediaries "to monitor the information which they transmit or store, nor a general obligation actively to seek facts or circumstances indicating illegal activity."[146] In addition, Article 10 of the European Convention on Human Rights also limits third-party liability for user-generated online content on free expression grounds.[147]

While service providers were not required to develop or deploy proactive monitoring or filtering techniques, they soon did so anyway. The scope and extent of copyright infringement made *ex post* notice and takedown a deeply unsatisfying remedy for copyright holders, who also urged legislators and regulators to require platforms to do more to stem the tide of infringing content.[148] In 1997, a number of commercial copyright owners and providers of user-generated content (UGC) services entered into

[142] E-Commerce Directive 2000/31, *supra* note 40, at art. 14.

[143] *See* Jennifer M. Urban et al., *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. Copyright Soc'y U.S.A. 371, 373 (2017).

[144] 17 U.S.C. § 512(m); E-Commerce Directive 2000/31, *supra* note 40, at art. 15.

[145] 17 U.S.C. § 512(m)(1). *Cf. id.* § 512(i)(2) (defining "standard technical measures" as measures that are "developed pursuant to a broad consensus of copyright owners and service providers in an open, fair, voluntary, multi-industry standards process; are available to any person on reasonable and nondiscriminatory terms; and do not impose substantial costs on service providers or substantial burdens on their systems or networks."). *See also* Lital Helman & Gideon Parchomovsky, *The Best Available Technology Standard*, 111 Colum. L. Rev. 1194, 1200 (2011) ("[S]ection 512 does not require webhosts to monitor content on their site *ex ante* as a prerequisite for enjoying the safe harbor.").

[146] E-Commerce Directive 2000/31, *supra* note 40, at art. 15.

[147] *See* Magyar Tartalomszolgáltatók Egyesülete v. Hungary, 2016 Eur. Ct. H.R. 1, 11-15 (finding that Article 10 of the European Convention on Human Rights (ECHR) limits third-party liability for online content); *but see* Delfi AS v. Estonia, 2013 Eur. Ct. H.R. 1, 33-34 (finding that imposing liability for unlawful, online hate speech does not offend Article 10 of the ECHR).

[148] Cohen, *supra* note 10, at 123-24 (describing the copyright wars). *See also* Nicolas P. Suzor, Lawless: The Secret Rules That Govern Our Digital Lives 76-78 (Cambridge Univ. Press 2019) (describing efforts to pass the Stop Online Piracy Act and the Preventing Real Online Threats to Economic Creativity and Theft of Intellectual Property Act, and require Internet platforms to do more to combat infringement).

the "UGC Principles," a non-binding set of principles that, among other things, called for UGC services to use "effective content identification technology" to eliminate infringing content from their services.[149]

YouTube's fingerprinting technology, Content ID, is a prime example of automated copyright enforcement.[150] Content ID, created in 2007, matches references submitted by copyright owners against user uploads to YouTube's website.[151] Using Content ID and similar fingerprint-based systems, rightsholders can opt to block infringing content in user-generated videos. Large rightsholders have developed automated mechanisms to detect, track, and report online infringement as well as to generate takedown requests.[152] And service providers have also adopted automated means to respond to notices of claimed infringement, thereby "automat[ing] the process in order to manage floods of requests."[153]

But Content ID illustrates that the dispute about proactive filtering of infringing content is not merely about deleting or blocking said content. Automated copyright enforcement has also enabled rightsholders to explore other remedies beyond takedown. For example, in addition to blocking content, Content ID also allows rightsholders to monetize that content—to redirect any revenue from the user who generated the video to the rightsholder—or to track the usage of that content.[154] The ability to monitor and monetize infringement remedies what rightsholders call the "value gap"[155] between what YouTube pays for monetized content and what services such as Spotify or Pandora, which license content directly from rightsholders, pay. Automated copyright enforcement provides a wealth of avenues for rightsholders to redress the "value gap" through means other than blocking access to content, for example, by monetizing or surveilling the usage or viewing patterns of content.[156]

[149] *Principles for User Generated Content Services*, UGC, https://ugcprinciples.com [https://perma.cc/V4XQ-PPXG] (last visited Apr. 16, 2020). *See also* Lauren G. Gallo, *The (Im)possibility of "Standard Technical Measures" for UGC Websites*, 34 Colum. J.L. & Arts 283, 295 (2011) (describing how "all major right holders and all major UGCs," even the ones that had declined to join the UGC Principles, had adopted some kind of content fingerprinting technology).

[150] Annemarie Bridy, *The Price of Closing the "Value Gap": How the Music Industry Hacked EU Copyright Reform*, 22 Vand. J. Ent. & Tech L. 323, 330 (2020) [hereinafter *The Price of Closing the "Value Gap"*].

[151] *Qualifying for Content ID*, YouTube, https://support.google.com/youtube/answer/1311402 [https://perma.cc/Y6XV-LQUN] (last visited Apr. 16, 2020).

[152] Niva Elkin-Koren, *Fair Use by Design*, 64 UCLA L. Rev. 1082, 1087 (2017); Urban et al., *supra* note 143, at 374.

[153] Elkin-Koren, *supra* note 152, at 1087.

[154] *See* Perel & Elkin-Koren, *supra* note 11, at 480, 510. *See also* Miguel Helft, *Google Told to Turn Over User Data of YouTube*, N.Y. Times (July 4, 2008), https://www.nytimes.com/2008/07/04/technology/04youtube.html [https://perma.cc/E6QN-D5JA] (describing how, in the Viacom contributory infringement case against YouTube, YouTube was required to produce data regarding user viewing histories).

[155] *See, e.g.*, *Medium: Five Stubborn Truths About YouTube and the Value Gap*, RIAA News (Aug. 18, 2017), https://www.riaa.com/medium-five-stubborn-truths-youtube-value-gap/ [https://perma.cc/ECV6-CH2X].

[156] *See, e.g.*, Audible Magic, https://www.audiblemagic.com/ [https://perma.cc/2C6M-XCB4] (last visited Apr. 16, 2020).

As technology has facilitated faster and farther-reaching takedown requests, many have cautioned that compliance with this framework may lead to over-deletion of lawful content. In particular, as one important study of notice and takedown found, "the rise of mass notice sending via automated systems raises immediate questions of accuracy and due process," because the sheer scale of automated notice sending makes it difficult to analyze the legal issues presented or understand whether notices are sent with bad faith.[157]

The architecture of Content ID, Audible Magic, and similar "fingerprinting" technologies also necessarily raises difficult questions about context. Fingerprinting techniques work by automatically screening user-generated content against an existing database of copyright-protected content: any clip of copyrighted material that matches protected content will be flagged as infringement.[158] The result is that, in their current form, automated systems for detecting copyright infringement are often incapable of detecting uses of copyrighted works that are non-infringing, including fair use.[159] The same is true for other exceptions that carve out other kinds of creative reuses of copyrighted materials, including "quotation, criticism, and review," or "caricature, parody, or pastiche."[160]

The need for an authoritative set of unlawful content, therefore, limits the applications of hash- or fingerprint-based technology. Consider the difficulty of using a fingerprinting approach to identify "hate speech." Compiling an authoritative set of "hate speech" would be impossible in its own right, and any effort to do so would be necessarily, indelibly influenced by political and social judgment. For instance, while it is a criminal offense under German law to display a swastika, the law recognizes several context-dependent exceptions, including "to promote art or science, research or teaching, reporting about current or historical events, or similar purposes."[161] A fingerprinting-based approach that identified photographs of swastikas as impermissible would also include journalism, art, and research in its sweep. It is nearly impossible to imagine such an approach being useful to platforms or to law enforcement.

While fingerprinting and hashing technologies are unlikely to helpfully address highly context-dependent questions, they are quite effective modes of public-private cooperation on policing content that is predetermined to be unlawful.[162] This explains their use in settings like child sex-

[157] Urban et al., *supra* note 143, at 406-09.

[158] *Core Technology & Services Overview*, AUDIBLE MAGIC (2015), https://www.audiblemagic.com/wp-content/uploads/2015/04/AM_overview_datasheet_150406.pdf [https://perma.cc/Z6AU-HN89] (describing "fingerprint" technology).

[159] *See Is Online Copyright Enforcement Scalable?*, *supra* note 11, at 715.

[160] *See Article 13 in 10 Questions*, ARTICLE 13, https://www.article13.org/faq [https://perma.cc/ULE4-6QU4] (last visited Apr. 16, 2020).

[161] STRAFGESETZBUCH [StGB] [PENAL CODE], § 86, para. 3, *translation at* https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html [https://perma.cc/BDU4-7PGM] (Ger.).

[162] *See, e.g.*, Martin, *supra* note 104, at 701 (describing how the National Software Reference Library provides hash sets for standard files to law enforcement to ease computer searches).

ual abuse imagery, in which a speaker is liable regardless of context.[163] But today, the content moderation decisions that platforms often seek to automate require more contextual analysis and human judgment than screening for unlawful content. Thus, platforms increasingly rely upon a combination of fingerprinting approaches, machine learning, and human decision-making to engage in content moderation. As large platforms set and enforce ground rules for membership, participation, and exclusion from online communities, they rely on a "bionic" combination of *ex ante* automated screening with *ex post* analysis by human moderators and algorithmic decision-making technologies.[164]

II. From Reactive to Proactive

Put simply, although the dominant mode of regulating unlawful content was reactive, platforms quickly developed their own proactive methodologies, transforming their ability to enforce private rules and lending their technological capabilities to the enforcement of offline laws. Today, however, the terms of these bargains are dramatically contested, as lawmakers in Europe and elsewhere consider and adopt requirements that platforms engage in *ex ante* monitoring and filtering.

In the discussion that follows, this Part maps new proactive monitoring requirements along several axes: the expanding role of private enforcement of technology and quasi-legal protections as instruments of private governance, the relationship between law enforcement agencies and private entities, the expansion of the kinds of content considered unlawful, and the dueling emphases on rapid takedowns and due process for restoring wrongfully deleted content. As platforms invest in artificial intelligence and algorithmic content moderation to counter the flood of toxic information, other government actors are seizing on their promises about the capacity of technology and asking platforms to do even more to proactively head off these threats.

A. Copyright

Article 17 of the EU Copyright Directive (Article 17), which was formerly known as Article 13, has fundamentally altered Europe's intermediary safe harbor protections.[165] Article 17 makes "online content-sharing service providers" liable when users upload copyright-infringing content

[163] *See* Gabriel J.X. Dance & Michael H. Keller, *How Laws Against Child Sexual Abuse Imagery Can Make It Harder to Detect*, N.Y. Times (Nov. 12, 2019), https://www.nytimes.com/2019/11/12/us/online-child-sex-abuse.html [https://perma.cc/4TQS-9HW9] (explaining how the ban on possessing or viewing child sexual abuse imagery has slowed down the private sector's development of new tools to detect it).

[164] *See Content Moderation: The Future is Bionic*, Accenture 1, 2 (2017), https://www.accenture.com/cz-en/_acnmedia/PDF-47/Accenture-Webscale-New-Content-Moderation-POV.pdf [https://perma.cc/RP2Q-W42W].

[165] Directive 2019/790, of the European Parliament and of the Council of 17 April 2019 on Copyright and Related Rights in the Digital Single Market and Amending Directives 96/9/EC and 2001/29/EC, art. 17, 2019 O.J. (L 130) 92, 120 [hereinafter Copyright Directive 2019/790].

unless the providers make their "best efforts" to license the content from the rights holder.[166] In addition, the new provisions require providers to "ensure the unavailability" of unlicensed content and to "expeditiously" remove and block future uploads of infringing content.[167] The provisions will apply to any provider that is over three years old, no matter how small, raising concerns that they will further entrench the dominance of existing platforms.[168]

Although Article 17 does not explicitly require proactive monitoring of user content—indeed, the provision states that it "shall not lead to any general monitoring obligation"—numerous critics have pointed out that it will nonetheless have the *de facto* effect of leading to proactive monitoring.[169] As German Member of European Parliament Julia Reda has described these provisions, service providers "will have no choice but to deploy upload filters" to block infringing content.[170]

Yet, despite strongly encouraging proactive, *ex ante* screening of user-generated content, Article 17 gives scant guidance to online content-sharing service providers regarding the technical methodologies of monitoring user-generated content or blocking infringement. Though it recognizes that automated, content screening will, by its very nature, broadly affect free expression, the Copyright Directive places no limitations on its design or use, instead putting its faith in platforms to develop mechanisms to review the individual challenges of blocking decisions.[171]

Rather than challenging the central position of major online platforms in the digital economy, Article 17 thus reaffirms it. This dynamic is partly a response to the logic of scale: the experience of Content ID already illustrates that, as platforms develop techniques for policing infringement, rightsholders demand broader and more extensive applications.[172] Allowing platforms to develop technical measures for compliance virtually ensures that *ex ante* content moderation minimizes the potential risk of liability.

Although proponents of Article 17 described the provision as enhancing "platform accountability," the broad discretion it confers on technology

[166] *Id.* at art. 17. *See also The Price of Closing the "Value Gap"*, *supra* note 150, at 357–58.

[167] Copyright Directive 2019/790, *supra* note 165, at art. 17(4).

[168] *The Price of Closing the "Value Gap"*, *supra* note 150, at 355–56.

[169] *See, e.g.*, *id.* at 353 ("the preventive measures demanded in the adopted text cannot realistically be achieved at scale without an [automated content recognition] system like Content ID."); Danny O'Brien, *EU's Parliament Signs Off on Disastrous Internet Law: What Happens Next?*, EFF (Mar. 26, 2019), https://www.eff.org/deeplinks/2019/03/eus-parliament-signs-disastrous-internet-law-what-happens-next [https://perma.cc/BA43-F6SL] (describing the inconsistency between Article 17 and the E-Commerce Directive).

[170] *The Text of Article 13 and the EU Copyright Directive Has Just Been Finalized*, Julia Reda (Feb. 13, 2019), https://juliareda.eu/2019/02/eu-copyright-final-text/ [https://perma.cc/DG2K-T6SN].

[171] *Id.*

[172] *See, e.g.*, *The Price of Closing the "Value Gap"*, *supra* note 150, at 108–11; Helft, *supra* note 154.

companies will, in the long term, serve only to further empower them.[173] Indeed, Article 17's approach exemplifies nearly unbridled deference to technology for addressing copyright infringement, as well as deference to the kinds of procedures that platforms believe safeguard due process.[174] Advocates of the provisions have deemphasized the significance of these changes, suggesting that because YouTube already employs Content ID to filter uploads, the new provisions will just be more of the same.[175]

Article 17's safeguards likewise shift the costs of protecting free expression to individual users.[176] Instead of requiring rightsholders to submit notices of claimed infringement, the Copyright Directive puts the onus on users to file a "counter-notice" demonstrating that their use of a copyrighted work falls into an exception or limitation to copyright protection.[177] While this mechanism formally pays lip service to the need to protect free expression, empirical studies have shown that counter-notice has served as an ineffective check on over-blocking.[178]

Not only does the Copyright Directive stress the need for expeditious takedowns, it also emphasizes the need for an "effective complaint and redress mechanism" in achieving the appropriate balance between free expression and copyright protection.[179] While it encourages platforms to use automated means to facilitate takedowns, Article 17(9) requires platforms to enact appeal mechanisms subject to "human review" without "undue delay." By creating a system in which takedowns are automated, but appeals are manual, Article 17 ensures that while takedowns occur at scale, appeals almost certainly cannot.

Although private-sector innovation may address some of the substantive concerns about over-blocking and other burdens on protected expression, the Copyright Directive also raises questions about competition.[180] By remaining silent on the design of algorithmic filtering, Article 17 does not create specific incentives for platforms to innovate in ways that pro-

[173] *See, e.g.*, Madigan, *supra* note 8 (describing online platforms as "the most powerful and wealthy entities in the world"); *Europe Takes an Important Step Toward Platform Accountability With Directive on Copyright*, AAP (Sept. 13, 2018), https://newsroom.publishers.org/europe-takes-an-important-step-toward-platform-accountability-with-directive-on-copyright/ [https://perma.cc/SW9D-CYZP].

[174] *Cf.* Madigan, *supra* note 8.

[175] Robert Levine, *Mind the Value Gap: Will Europe Address the Legal Loophole That Lets YouTube Pay Less for Music?*, BILLBOARD (Sept. 11, 2018), https://www.billboard.com/articles/business/8474670/mind-value-gap-europe-address-legal-loophole-lets-youtube-pay-less-music-column [https://perma.cc/M9NG-FABU].

[176] *See generally* Urban et al., *supra* note 143.

[177] *Id.* 393–94, 405.

[178] Elkin-Koren, *supra* note 152, at 1091–92; Urban et al., *supra* note 143, at 406–10.

[179] *See* Copyright Directive 2019/790, *supra* 165, at 108.

[180] *See* Elkin-Koren, *supra* note 152, at 1097 (suggesting that improvements to filtering technology might mitigate some of the concerns that proactive screening sweeps too broadly: as automated copyright enforcement moves toward artificial intelligence and machine learning, platforms may be able to design systems that are friendlier to fair uses and that "learn patterns of fair use instances by studying existing fair use decisions").

mote fair use.[181] Moreover, many of the companies that have already developed proactive filtering and blocking software stand to benefit enormously from the uptick in new customers. Indeed, the German Data Protection Commissioner has raised concerns that, because Article 17 all but requires upload filters, small companies will rely on the filtering technologies of larger platforms, leading to the emergence of an "oligopoly" of filtering software.[182]

B. Unlawful Speech

Despite the fact that most of the large platforms are now using proactive automated means to filter and block terrorist content on a purportedly voluntary basis, the availability and widespread positive publicity about automated, *ex ante* monitoring and blocking of extremist content has prompted multiple government actors to enact, or consider, legislation that would make this technology virtually compulsory. In spring 2019, Australia enacted a law that will impose significant penalties on online service providers if they fail to rapidly remove "abhorrent violent material" from their services.[183] Under the new statute, providers of online "content services" commit a criminal offense if their services host "abhorrent violent material" that they fail to "expeditiously" remove.[184] The statute does not define "expeditiously," but the Australian Attorney General expressed, in a reading speech, his conviction that platforms could address these concerns through the use of technology.[185] The statute applies regardless of the size of the company.[186]

Germany's new Network Enforcement Act of 2018, colloquially known as NetzDG, similarly imposes burdens on social media platforms, and these burdens virtually require the use of upload filters. Under

[181] Stan Adams, *Why the EU Copyright Directive Is a Threat to Fair Use*, CTR. DEMOCRACY & TECH. (Mar. 1, 2019), https://cdt.org/blog/why-the-eu-copyright-directive-is-a-threat-to-fair-use/ [https://perma.cc/4LDJ-ACHG].

[182] Press Release, Copyright Reform Also Harbors Data Protection Risks, BFDI (Feb. 26, 2019) (available at https://translate.google.com/translate?hl=en&sl=de&u=https://www.bfdi.bund.de/DE/Infothek/Pressemitteilungen/2019/10_Uploadfilter.html&prev=search&pto=aue [https://perma.cc/BW9U-V2RD]) [hereinafter BFDI Press Release] ("Ultimately, such an oligopoly would result in fewer providers of filter technologies, through which more or less all of the internet traffic of relevant platforms and services runs.").

[183] Damien Cave, *Australia Passes Law to Punish Social Media Companies for Violent Posts*, N.Y. TIMES (Apr. 3, 2019), https://www.nytimes.com/2019/04/03/world/australia/social-media-law.html [https://perma.cc/5JPH-VL5G].

[184] *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019* (Cth) s 474.34 (Austl.).

[185] Evelyn Douek, *Australia's "Abhorrent Violent Material" Law: Shouting "Nerd Harder" And Drowning Out Speech*, 94 AUSTRALIAN L.J. 41 (forthcoming 2020) [hereinafter *Australia's "Abhorrent Violent Material" Law*]. *See also* Daphne Keller, *Three Constitutional Thickets: Why Regulating Online Violent Extremism Is Hard*, GEORGE WASH. PROGRAM ON EXTREMISM 1, 3 (2019), https://extremism.gwu.edu/sites/g/files/zaxdzs2191/f/Three%20Constitutional%20Thickets.pdf [https://perma.cc/J48E-URDT] [hereinafter *Three Constitutional Thickets*] (summarizing the current status of several statutes pushing platforms to take down extreme content more quickly).

[186] *See Australia's "Abhorrent Violent Material" Law*, *supra* note 185.

NetzDG, platforms must remove certain kinds of "manifestly unlawful" speech—defined by reference to the German criminal code—within twenty-four hours or face heavy penalties.[187] NetzDG applies to social network providers with two million, or more, registered users in Germany.[188]

The EU's draft regulation on preventing the dissemination of terrorist content online likewise requires platforms to remove or disable access to "terrorist content" within one hour, or face heavy fines.[189] The draft regulation also adopts a relatively broad definition of "terrorist content," which includes not only direct incitement but also "glorifying" terrorist crimes or "depicting the commission" of a terrorist offense.[190]

Like Article 17, the AVM law, NetzDG, and the EU's draft terrorism regulation do not overtly require platforms to adopt proactive screening methodologies. Nevertheless, their efforts to scale enforcement of the law push them in that direction. Google's NetzDG transparency report, for instance, documents how it uses hashing, fingerprinting, and automated flagging technologies to try to identify unlawful content more quickly.[191]

Yet it is particularly difficult to automate compliance with these kinds of provisions because determining whether speech is unlawful depends on the context. For instance, NetzDG reaches far more broadly than "terrorist content" and applies to unlawful content that includes "public incitement to commit offences" and "disturbing public peace by threatening to commit offences."[192] The difficulty of conducting complex, fact-dependent analysis of whether material is, in fact, unlawful helps to explain why, as Google notes, "[m]achine automation simply cannot replace human judgment and nuance."[193]

This difficulty actually prompted the European Parliament to abandon an earlier effort that required platforms to develop new proactive technologies of automated content moderation, and instead endorse a narrower approach of requiring "specific" measures short of proactive monitoring.[194] Under an earlier version of the draft terrorism regulation, these measures could include automated removal of content, automated preven-

[187] Evelyn Douek, *Germany's Bold Gambit to Prevent Online Hate Crimes and Fake News Takes Effect*, LAWFARE (Oct. 31, 2017, 11:30 AM), https://www.lawfareblog.com/germanys-bold-gambit-prevent-online-hate-crimes-and-fake-news-takes-effect [https://perma.cc/V9ZU-BRN3].

[188] *Id.* (noting that "registered" users is more inclusive than "active" users).

[189] *European Parliament Legislative Resolution of 17 April 2019 on the Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*, at art. 4(2), COM (2018) P8_TA(2019)0421 (Sept. 4, 2019) [hereinafter *Draft Terrorism Regulation*].

[190] *Id.* at art. 2.

[191] *Removals Under the Network Enforcement Law*, GOOGLE, https://transparencyreport.google.com/netzdg/youtube?hl=en [https://perma.cc/SF24-X6ZK] (last visited Feb. 5, 2020) [hereinafter *Removals*].

[192] STRAFGESETZBUCH [StGB] [PENAL CODE], §§ 111, 126, *translation at* https://www.gesetze-im-internet.de/englisch_stgb/englisch_stgb.html [https://perma.cc/BDU4-7PGM] (Ger.).

[193] *Removals*, *supra* note 191.

[194] *See Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*, at art. 6, COM (2018) 640 final

tion of re-uploading, or "detecting, identifying and expeditiously removing" new terrorist content.[195] The push toward automation responded directly to fears that "the longer the content is able to survive online, the more views it may receive, and the more harm it may cause."[196] Strikingly, the European Commission had argued that requiring platforms, on pain of liability, to develop technologies to filter and monitor content was still consistent with the E-Commerce Directive's immunity provision.[197]

The European Parliament's version of the draft terrorism regulation is also more sensitive to the risk that these obligations might tend to entrench dominant platforms. The regulation instructs that any request for "specific measures" that platforms ought to take should account for the "technical feasibility of the measures, the size and economic capacity" of the platform, and the effects of the measures on free expression and the free flow of information.[198] The draft regulation also makes clear that any penalties, including fines, should account for the "financial resources" of the platform, whether the platform is a start-up or a small- to medium-sized business, and whether it could comply with a removal order.[199]

Platforms have generally opposed laws that impose these kinds of obligations. Facebook vigorously opposed NetzDG, for example, arguing that the state was "pass[ing] on its own shortcomings and responsibilities to private companies."[200] Similarly, a consortium of technology companies opposed the Australian AVM measure, arguing that the government did not consult with the technology sector before drafting the bill.[201]

Whether required by law or not, however, platforms are in fact committing to develop proactive, automated screening methodologies in

(Sept. 12, 2018) [hereinafter *Commission Draft Terrorism Regulation*] (calling for proactive monitoring).

[195] *Id.*

[196] *Commission Staff Working Document Impact Assessment*, at 13, SWD (2018) 408 final (Sept. 12, 2018).

[197] *See Commission Draft Terrorism Regulation*, *supra* note 194, at § 1.2 ("The present proposal is consistent with the acquis related to the Digital Single Market and in particular the E-Commerce Directive.").

[198] *Draft Terrorism Regulation*, *supra* note 189, at art.6. *See also Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision*, CRE 1, 17 (2019), http://thecre.com/RegSM/wp-content/uploads/2019/05/French-Framework-for-Social-Media-Platforms.pdf [https://perma.cc/V7S7-HFJQ] (suggesting a sliding scale of compliance obligations depending on the size of the operator and its service).

[199] *Draft Terrorism Regulation*, *supra* note 189, ¶ 38. When this Article was finalized in April 2020, the draft regulation was undergoing trialogue negotiations. Leaked documents suggested that these Parliamentary changes were not going to be included in the final text. *See EU Online Terrorist Content Legislation Risks Undermining Press Freedom*, CPJ (Mar. 11, 2020, 6:46 AM), https://cpj.org/2020/03/eu-online-terrorist-content-legislation-press-freedom.php [https://perma.cc/D49P-JWF2]. The leaked documents suggested that the final version of the regulation would not require automated filtering. *Id.*

[200] *See* Jefferson Chase, *Facebook Slams Proposed German 'Anti-Hate Speech' Social Media Law*, DEUTSCHE WELLE (May 29, 2017), https://www.dw.com/en/facebook-slams-proposed-german-anti-hate-speech-social-media-law/a-39021094 [https://perma.cc/5JCR-227L].

[201] *See* Cave, *supra* note 183.

response to political pressure as well as legislation. In particular, fingerprinting and hash-based screening are also being touted as code-based solutions to the problems of online hate speech and terrorist speech.[202] Nevertheless, as NetzDG illustrates, local laws are prompting platforms to develop proactive screening methodologies for a host of other types of illicit content and to invest in expanding *ex post* moderation capabilities.

C. Defamation

A third context in which regulators have compelled platforms to use automated means to filter or block online speech occurs in the context of court judgments ordering online platforms to screen user-generated content for future instances of illegality. In October 2019, the European Court of Justice (CJEU) issued an opinion partially upholding an injunction that required Facebook to delete a comment calling Austrian politician Eva Glawischnig-Pieszek a "corrupt oaf" because it was defamatory under Austrian law.[203] At the core of Glawischnig-Pieszek's case was her request for an injunction that would compel Facebook to delete any identical or "equivalent" statements posted by any user worldwide.[204] Article 15 of the E-Commerce Directive precludes EU member states from imposing "general obligations to monitor" the information transmitted or stored by hosting providers.[205] In a previous case, *SABAM v. Netlog*, the CJEU had held that a Belgian court could not issue an injunction that "requires [a host] to install a system for filtering" specific types of content "indiscriminately" as to all of its users.[206]

Nonetheless, the CJEU concluded that the Austrian court could, consistent with Article 15, order Facebook to remove defamatory material, identical reposts, and "information, the content of which, whilst essentially conveying the same message, is worded slightly differently."[207] At the root

[202] *See, e.g.*, Sissi Cao, *Facebook's AI Chief Explains How Algorithms Are Policing Content–And Whether It Works*, OBSERVER (Dec. 6, 2019, 7:15 AM), https://observer.com/2019/12/facebook-artificial-intelligence-chief-explain-content-moderation-policy-limitation/ [https://perma.cc/4V29-QHAL] ("And, in addition to that, we have a number of AI tools that we are developing, like the ones that I had mentioned, that can proactively go flag the content.").

[203] Case C–18/18, Glawischnig-Piesczek v. Facebook Ir. Ltd., 2019 E.C.R. ¶ 53. *See also* DAPHNE KELLER, DOLPHINS IN THE NET: INTERNET CONTENT FILTERS AND THE ADVOCATE GENERAL'S *GLASWISCHNIG-PIESCZEK V. FACEBOOK IRELAND* OPINION 2 (Stanford Ctr. for Internet & Soc'y 2019).

[204] *Glawischnig-Piesczek*, 2019 E.C.R. ¶ 20.

[205] E-Commerce Directive 2000/31, *supra* note 40, at art. 15. *See also* Aleksandra Kuczerawy, *To Monitor or Not to Monitor? The Uncertain Future of Article 15 of the E-Commerce Directive*, BALKINIZATION (May 29, 2019), https://balkin.blogspot.com/2019/05/to-monitor-or-not-to-monitor-uncertain.html [https://perma.cc/4ACH-B4JM].

[206] Case C-360/10, SABAM v. Netlog, 2012 E.C.R. ¶ 26. *See also* Case C-484/14, Mc Fadden v. Sony Music Entm't Ger. GmbH, 2016 E.C.R. ¶ 87; Case C-324/09, L'Oréal v. eBay, 2011 E.C.R. ¶ 139 ("The measures required of the online service provider concerned cannot consist in an active monitoring of all the data of each of its customers in order to prevent any future infringement of intellectual property rights via that provider's website.").

[207] *Glawischnig-Piesczek*, 2019 E.C.R. ¶ 41.

of the CJEU ruling is the conviction that Facebook can use "automated search tools and technologies" to identify new posts of offending material.[208] Resting on this assumption, the CJEU held that the injunction was permissible so long as it identified "specific elements . . . such as . . . equivalent content to that which was declared to be illegal," and so long as Facebook was not required to conduct an "independent assessment" of whether content was covered or not.[209]

The opinion neglected to examine several critical aspects of the relief Glawischnig-Pieszek sought, however. First, algorithmic filtering of defamatory statements is difficult because it requires context: how could automated methods, for instance, tell the difference between a defamatory comment and a news report on the case?[210] As Jennifer Daskal and Kate Klonick pointed out before the opinion came down, "it's much more difficult than it sounds to define, let alone reliably identify, an 'identical' post."[211] Moreover, the record of automated search tools and technologies that Facebook has at hand to conduct such monitoring, blocking, and filtering is scant. In particular, the CJEU offered no evidence to support its impression that the "search tools and technologies" that Facebook has access to neither required "independent assessment" nor constituted "indiscriminate" filtering.[212]

By suggesting that the availability of "automated search tools and technologies" minimized the role Facebook had to play in determining whether content was within the scope of the injunction, the CJEU expressed—whether intentionally or not—a latent trust in the capacity of automated systems to make judgments about content.[213] In so doing, the CJEU seemed to rely on Advocate General Maciej Szpunar's conclusion that requiring Facebook to monitor and delete reposting of identical statements was a proportionate remedy because "seeking and identifying information identical to that which has been characterized as illegal by a court . . . does not require sophisticated techniques that might represent an extraordinary burden."[214] Instead, the Advocate General suggested that "identical" statements could be detected "with the help of software tools."[215]

In light of the scant factual record, it is difficult to understand Advocate General Szpunar's confidence in Facebook's technology, much less the

[208] *Id.* ¶ 46.

[209] *Id.* ¶ 45.

[210] *See generally* KELLER, *supra* note 203 (analyzing the potential over inclusiveness of the approach urged by the Advocate General). *Cf.* James Vincent, *Zuckerberg Criticized over Censorship After Facebook Deletes 'Napalm Girl' Photo*, VERGE (Sept. 9, 2016, 5:18 AM), https://www.theverge.com/2016/9/9/12859686/facebook-censorship-napalm-girl-aftenposten [https://perma.cc/6DZW-8JUS].

[211] Jennifer Daskal & Kate Klonick, *When a Politician Is Called a 'Lousy Traitor,' Should Facebook Censor It?*, N.Y. TIMES (June 27, 2019), https://www.nytimes.com/2019/06/27/opinion/facebook-censorship-speech-law.html [https://perma.cc/C4A2-TQ2W].

[212] *Glawischnig-Piesczek*, 2019 E.C.R. ¶ 46.

[213] *Id.* ¶ 45.

[214] *Id.* ¶ 87 (separate opinion of Advocate General Maciej Szpunar).

[215] *Id.* ¶ 61.

CJEU's determination that automated solutions could also discern content *equivalent*—but not identical—to defamation.[216] But one might speculate that the widespread publicity about automated content moderation might seem to support the CJEU's findings. Ironically, although platforms themselves have "trumpet[ed] the technologies' capabilities" to avoid regulation, the CJEU may have had those promises in mind when it found that Facebook could use software magic to prevent the republication of defamatory content.[217]

In a sense, the Glawischnig-Piesczek case is nothing new. Courts have been fighting about the appropriate scope of injunctive relief for defamation for decades.[218] Injunctions against online intermediaries are uniquely effective methods of addressing defamation and other kinds of harmful speech, which is precisely why intermediary liability protections are so important.[219] In that respect, there have been ongoing debates under European law about the boundary between appropriate injunctive relief on the one hand, and the unlawful imposition of a general monitoring obligation on the other.[220] In the CJEU's view, requiring Facebook to monitor and delete future posts that were identical or equivalent to those judged defamatory was well within this boundary. However, by tasking Facebook with using automated methods to detect content equivalent to defamation, the CJEU is in novel territory. It is easy to imagine how the compliance technologies Facebook might use could be easily transferred to other contexts or settings.

III. The Drawbacks of Proactive Moderation

Governments increasingly view automated, content moderation as an appealing mechanism for solving the full range of "bad content" problems on social media. Drawing on the apparent success of algorithmic copyright enforcement, automation is now being touted as a solution to the

[216] *Id.* ¶ 73.

[217] *Internet Platforms*, *supra* note 114, at 7.

[218] *See, e.g.*, David S. Ardia, *Freedom of Speech, Defamation, and Injunctions*, 55 Wm. & Mary L. Rev. 1, 6 (2013); Steve Tensmeyer, *Constitutionalizing Equity: Consequences of Broadly Interpreting the "Modern Rule" of Injunctions Against Defamation*, 72 N.Y.U. Ann. Surv. Am. L. 43, 44 (2017). *See generally* Vincent Blasi, *Toward a Theory of Prior Restraint: The Central Linkage*, 66 Minn. L. Rev. 11 (1981); Erwin Chemerinsky, *Injunctions in Defamation Cases*, 57 Syracuse L. Rev. 157 (2007); John Calvin Jeffries, *Rethinking Prior Restraint*, 92 Yale L.J. 409 (1983).

[219] *See, e.g.*, Paul Schiff Berman, *Legal Jurisdiction and the Deterritorialization of Data*, 71 Vand. L. Rev. 11, 27–28 (2018) ("Governments have always enacted regulation through powerful intermediaries."). *See also* Hassell v. Bird, 420 P.3d 776, 789–90 (Cal. 2018); Martin Husovec, Injunctions Against Intermediaries in the European Union: Accountable but Not Liable? 57 (Cambridge Univ. Press 2017); Eleonora Rosati, *Intermediaries and IP: 5 Key Principles of EU Law*, IPKat (May 21, 2018), http://ipkitten.blogspot.com/2018/05/intermediaries-and-ip-5-key-principles.html [https://perma.cc/H9UT-7HDJ].

[220] Keller, *supra* note 203, at 29–31; Eleonora Rosati, *Material, Personal and Geographic Scope of Online Intermediaries' Removal Obligations Beyond* Glawischnig-Piesczek, C-18/18 *and Defamation*, Eur. Intell. Prop. Rev. (forthcoming), *available at* https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3438102 [https://perma.cc/NZ26-3TAJ].

problems of defamation, terrorist content, and other harmful speech online.[221] This Part explores how requiring platforms to deploy automated means to restrict speech raises substantial concerns about collateral censorship, surveillance, algorithmic control, and private power.[222]

A. Content Moderation as Censorship

While early cyber enthusiasts predicted that the Internet would be a new world free of surveillance and speech regulation, intermediary immunity, in fact, has not eliminated censorship and monitoring, but rather privatized them.[223] Despite being nominally free of liability, platforms have proven to be vulnerable to what Jack Balkin calls "new-school" methods of speech regulation: "collateral censorship," "public-private cooperation and cooptation," and "digital prior restraint."[224] When states coerce platforms to cooperate in censorship and surveillance, they play into the central dynamic that characterizes the uneasy relationship amongst users that depend on platforms for effective communication, platforms that depend on governments for a favorable regulatory environment, and governments that depend on platforms to carry out vital law enforcement tasks.[225]

Ex ante, automated content moderation aptly illustrates this dynamic. Calls to extend *ex ante*, automated content moderation to particular types or categories of speech create the risk of collateral censorship and digital prior restraint by threatening to hold platforms liable unless they censor speech at the government's bidding.[226] Moreover, the companies that control the "infrastructure of free expression" provide only weak protections when a government "uses that infrastructure, or its limitations, as leverage for regulation or surveillance."[227]

As it stands, automated content moderation already demonstrates the risk that technical "solutions" designed to prevent bad content from spreading will have collateral effects on lawful expression. One recent, quantitative analysis of a random sample of over 1,800 DMCA takedown requests, found that a significant number of requests either, incorrectly identified, or insufficiently specified, the allegedly infringing work.[228] Despite these

[221] *See* discussion *supra* Part II.

[222] Portions of this Part draw on my previous article, *see generally* Bloch-Wehba, *supra* note 21.

[223] *Free Speech Is a Triangle*, *supra* note 10, at 2011–15; *Old-School/New-School Speech Regulation*, *supra* note 12, at 2298–99. *See also* Derek E. Bambauer, *Against Jawboning*, 100 MINN. L. REV. 51, 57–58 (2015) [hereinafter *Against Jawboning*]; Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 11 (2006).

[224] *Free Speech Is a Triangle*, *supra* note 10, at 2011.

[225] *Id.* at 2020, 2035, 2047.

[226] *Id.* at 2018–19 ("Imposing liability on infrastructure providers unless they surveil and block speech, or remove speech that others complain about, has many features of a prior restraint, although technically it is not identical to a classic prior restraint.").

[227] *Old-School/New-School Speech Regulation*, *supra* note 12, at 2303.

[228] URBAN ET AL., NOTICE AND TAKEDOWN IN EVERYDAY PRACTICE 2 (2d ed. 2017). *See also* Daphne Keller, *Empirical Evidence of "Over-Removal" by Internet Companies Under*

insufficiencies, material that is alleged to infringe a claimant's copyright is routinely "removed before the target [of a takedown request] is given the opportunity to respond."[229] Users who are targeted by a wrongful takedown request rarely send counter-notices, and the "unbalanced liability standards" of copyright make it legally risky for platforms to encourage their users to send counter-notices.[230] The result is a regime in which the technical and legal infrastructure for DMCA compliance appears to have come at a significant cost to users' interests in free expression.

These trade-offs are not, of course, unique to copyright enforcement. Take, for example, the Global Internet Forum to Counter Terrorism's hash-sharing database. The chief virtue of the hash database is its efficiency: because the database is shared across platforms, it prevents users from effectively re-uploading videos and images that have already been identified as violent.[231] But this efficiency comes at a substantial cost to free expression. Like automated copyright enforcement, the hash database for violent extremist and terrorist content is "context-blind"—as Daphne Keller has put it, "an ISIS video looks the same, whether used in recruiting or in news reporting."[232] The result is that the hash database may have a disproportionately negative effect on news organizations, human rights defenders, and dissidents who seek to expose and comment on violence.[233]

Platforms' efforts to proactively monitor and block user expression raise three particular concerns about collateral censorship. First, as a matter of substance, efforts to exclude certain categories of expression from public discourse are likely to target marginalized perspectives and underrepresented communities.[234] Nowhere is this more evident than in efforts to define "terrorist" and "extremist" speech. When governments pressure platforms to more aggressively address terrorist or extremist content online, they often reflect the teachings of years of Islamophobic security policy. As Amna Akbar has documented, the "discursive construct" of "radicalization" has irrevocably shaped government's identification of the "terrorist threat."[235] As a result, many mundane aspects of daily life in Muslim communities are understood by law enforcement as potentially significant indicators of "radicalism."[236] In turn, the technological infrastruc-

Intermediary Liability Laws, Stan. Ctr. for Internet & Soc'y (Oct. 12, 2015), http://cyberlaw.stanford.edu/blog/2015/10/empirical-evidence-over-removal-internet-companies-under-intermediary-liability-laws [https://perma.cc/383C-6VAH] (summarizing several studies).

[229] Urban et al., *supra* note 228, at 118.

[230] *Id.*

[231] *GIFCT Transparency Report*, GIFCT, https://www.gifct.org/transparency/ [https://perma.cc/S7L4-5KE4] (last visited July 7, 2020).

[232] *Internet Platforms*, *supra* note 114, at 7.

[233] Land, *supra* note 10, at 7, 60–61.

[234] *See* Sap et al., *supra* note 96, at 1668 (reviewing Perspective API's disproportionate flagging of African-American English speech as "toxic" and "rude").

[235] Amna Akbar, *Policing "Radicalization"*, 3 U.C. Irvine L. Rev. 809, 816 (2013).

[236] *Id.* at 827. *See also* Amna Akbar, *National Security's Broken Windows*, 62 UCLA L. Rev. 834, 834 (2015); Sahar F. Aziz, *Losing the "War of Ideas:" A Critique of Counter Violent Extremism Programs*, 52 Tex. Int'l L.J. 255, 257–58 (2017).

tures that platforms then build, reflect government's security priorities. Therefore, systems of automated content moderation build on social and political constructions of the terrorist threat—but which constructions, and at whose expense?

Consider, for example, platforms' inconsistent approaches to the problem of white nationalism, Nazism, and right-wing terrorism on the one hand, and to ISIS and al-Qaeda on the other. Though major platforms took steps after Christchurch to eliminate white nationalism and separatism from their services, it appears that they still struggle with the issue. An ongoing civil rights audit of Facebook points out that the company's ban extends only to content that explicitly uses the terms "white nationalism" or "white separatism," and thus "leaves up content that expressly espouses white nationalist ideology" without using these keywords.[237] In August 2019, the Anti-Defamation League published a list of white nationalist groups still active on YouTube after it famously "purged" them.[238] YouTube responded by removing some of the channels highlighted in the report, but many were able to restore access to the platform.[239]

These examples highlight the possibility that platforms are applying their own content moderation rules in ways that are, if not outright discriminatory, at the very least, underinclusive. Moreover, platforms highlighted the "effectiveness" of their filtering and blocking techniques as a signal to European and American legislators that they took the threat of Islamic terrorism seriously.[240] But only after Christchurch, as political pressure to address white nationalist terrorism ramped up, did online platforms begin to actually take this threat seriously. Even now, Facebook's ongoing public relations campaign to counter accusations that the platform discriminates against conservative viewpoints undermines its efforts to address white nationalism.[241]

Second, as a matter of process, platforms and governments are also willing to tolerate higher error costs for speech that is identified as a prior-

[237] *Facebook's Civil Rights Audit—Progress Report*, Facebook 1, 9 (2019), https://about.fb.com/wp-content/uploads/2019/06/civilrightaudit_final.pdf [https://perma.cc/T8A5-QUX2] [hereinafter *Facebook's Civil Rights Audit*].

[238] *Despite YouTube Policy Update, Anti-Semitic, White Supremacist Channels Remain*, ADL (Aug. 15, 2019), https://www.adl.org/blog/despite-youtube-policy-update-anti-semitic-white-supremacist-channels-remain [https://perma.cc/82SV-J9RN].

[239] *Id. See also* Alex Kaplan, *YouTube Removed Some Channels Affiliated with White Nationalism—But Not All*, Media Matters (Aug. 28, 2019, 2:26 PM), https://www.mediamatters.org/white-nationalism/youtube-removed-some-channels-affiliated-white-nationalism-not-all [https://perma.cc/T64E-82KU].

[240] *See* Bloch-Wehba, *supra* note 21, at 44–45 (documenting how EU policy evolved in response to a series of ISIS attacks in France).

[241] *See* Charlie Warzel, *Why Will Breitbart Be Included in 'Facebook News'?*, N.Y. Times (Oct. 25, 2019), https://www.nytimes.com/2019/10/25/opinion/mark-zuckerberg-facebook.html [https://perma.cc/BK73-6XZL]; Jason Wilson, *Leaked Emails Reveal Trump Aide Stephen Miller's White Nationalist Views*, Guardian (Nov. 14, 2019, 2:00 PM), https://www.theguardian.com/us-news/2019/nov/14/stephen-miller-leaked-emails-white-nationalism-trump [https://perma.cc/CW8K-FKS4].

ity for removal.[242] For instance, YouTube's recent "purge" of videos supporting white supremacy and white nationalism also swept up videos that documented, reported on, and aimed to counteract those ideologies.[243] Experience has shown that systems designed to block certain kinds of speech are likely to be overinclusive—hardly a novel observation.[244] In theory, overinclusiveness is a design flaw that can be overcome by technological innovation.[245] But accepting high error rates within systems designed to monitor, block, filter, and monetize user expression is a political decision. In practice, the incentives for platforms to take "bad content" down always seem to outweigh the incentives to design systems with minimal error rates or maximal accommodations for free speech.[246] These political pressures influence informational filters, although they are rarely accounted for by designers. Indeed, as companies find themselves scrutinized from all sides, political and social pressure will likely inform speech decisions as much as, or more than, technology alone.

Third, the new wave of Internet regulation and the emergence of "voluntary" filtering illustrates the risk that governments will *informally* pressure platforms to adopt limitations on speech (what Derek Bambauer has called "jawboning").[247] When governments do this through political means—for example, through formal regulation or legislation—that political act is formally accountable to the public. But when governments pressure platforms to use their private authority to take down certain types of speech—for example, because it violates their terms of service—they tend to do so in ways that are less visible and less accountable to the public.[248]

[242] *Cf. Internet Platforms*, *supra* note 114, at 7 (arguing that platforms with less money to invest in content recognition software "will, if forced to build filters, presumably be forced to tolerate high rates of false positives in order to avoid liability for false negatives").

[243] Kyle Daly, *YouTube to Ban Supremacist Content, Purge Videos*, POLITICO (June 5, 2019, 2:41 PM), https://politi.co/2HX3Jf2 [https://perma.cc/2RXZ-V739]; Kelly Weill, *YouTube Crackdown on Extremism Also Deleted Videos Combating Extremism*, DAILY BEAST (June 6, 2019, 3:20 PM), https://www.thedailybeast.com/youtube-crackdown-on-extremism-also-deleted-videos-combating-extremism [https://perma.cc/5CCA-78B7].

[244] *See, e.g.*, Duarte et al., *supra* note 97, at 6. *See also* EVAN ENGSTROM & NICK FEAMSTER, THE LIMITS OF FILTERING: A LOOK AT THE FUNCTIONALITY & SHORTCOMINGS OF CONTENT DETECTION TOOLS 19 (2017); J.M. Balkin, Comment, *Media Filters, the V-Chip, and the Foundations of Broadcast Regulation*, 45 DUKE L.J. 1131, 1133 (1996) [hereinafter *Media Filters*]; Derek Bambauer, *Cybersieves*, 59 DUKE L.J. 377, 397 (2009) [hereinafter *Cybersieves*]; *Internet Platforms*, *supra* note 114, at 6 (describing errors in automated generation of DMCA takedown requests, and noting that systems for processing those requests sometimes, but not always, catch errors); Land, *supra* note 10, at 8.

[245] Elkin-Koren, *supra* note 152, at 1097; *Media Filters*, *supra* note 244, at 1153–54 (noting that two chief arguments against the V-Chip—that "parents will be unable to use [it] . . . [and that] children will be able to break through. . . ."—are really a "problem of technological design").

[246] *See, e.g.*, Bloch-Wehba, *supra* note 21, at 79 (describing the "lopsided incentives" to remove online content).

[247] *Against Jawboning*, *supra* note 223, at 57.

[248] Chang, *supra* note 132, at 120–22; Land, *supra* note 10, at 15.

B. Content Moderation as Surveillance

Automated content moderation also opens up new avenues for surveillance and monitoring of users as individuals and as groups.[249] In this respect, hash- or fingerprint-based technologies like Content ID and the GIFCT hash database are particularly troubling. The GIFCT hash database offers a ripe, new way for platforms to identify, cross-reference, and keep tabs on accounts that have posted terrorist content in the past. Platforms have claimed that the database includes around 100,000 hashed images.[250] Yet, it is unclear whether, and under what circumstances, those images are shared with law enforcement.

The hash database has significant potential as a counterterrorist tool far beyond taking down terrorist content. Just as Content ID permits rightsholders to opt to monitor viewing activity on infringing videos, law enforcement might opt to monitor engagement with terrorist posts on social media in order to map associations and networks of suspected sympathizers, to understand the diffusion of propaganda, or simply to monitor those who have viewed dangerous online content.[251] Indeed, some have raised concerns that systematically deleting terrorist content from major platforms will drive terrorist networks underground, thereby depriving government of a critical source of information.[252]

Even apart from direct information sharing between GIFCT and the government, automated mechanisms illustrate how closely connected speech rights and surveillance are. To understand how private, automated moderation mechanisms might play into existing law enforcement paradigms, imagine that YouTube's algorithm for screening extremist content prevents a user from uploading a television segment reporting on the Kurdistan Worker's Party, or the PKK. Now imagine that a law enforcement

[249] *See, e.g.*, Kate Crawford & Jason Schultz, *Big Data and Due Process: Toward a Framework to Redress Predictive Privacy Harms*, 55 B.C. L. REV. 93, 94–95 (2014). The use of machine learning for algorithmic content moderation also has more concrete privacy harms beyond the context of law enforcement surveillance. For example, to the extent that platforms use machine learning methods which learn from user-generated content without the consent or knowledge of the users themselves, it might have significant privacy implications. *Id.*

[250] Larry Greenemeier, *Social Media's Stepped-Up Crackdown on Terrorists Still Falls Short*, SCI. AM. (July 24, 2018), https://www.scientificamerican.com/article/social-medias-stepped-up-crackdown-on-terrorists-still-falls-short/ [https://perma.cc/UEK4-XJUG].

[251] *See, e.g.*, Hannah Bloch-Wehba, *Process Without Procedure: National Security Letters and First Amendment Rights*, 49 SUFFOLK U. L. REV. 367, 381 (2016) (illustrating how national security letters are used to find associations and networks of individuals and groups).

[252] *See, e.g.*, Bateman, *supra* note 25. Strikingly, the American Civil Liberties Union (ACLU) made a similar argument in defense of its decision to represent white nationalist protestors in advance of the Charlottesville "Unite the Right" rally in 2017. *See* Anthony D. Romero, *Equality, Justice and the First Amendment*, ACLU (Aug. 15, 2017, 6:00 PM), https://www.aclu.org/blog/free-speech/equality-justice-and-first-amendment [https://perma.cc/HU3L-CVKR] ("Racism and bigotry will not be eradicated if we merely force them underground."). *See also* P.E. MOSKOWITZ, THE CASE AGAINST FREE SPEECH: THE FIRST AMENDMENT, FASCISM, AND THE FUTURE 18 (Bold Type Books 1st ed. 2019).

agency subpoenas YouTube for subscriber information pertaining to all users who have attempted to upload prohibited "terrorist" videos, among which our user is one.

In several respects, this hypothetical illustrates how content moderation rules—and the technical infrastructures that enforce them—might open new kinds of behavior and new actors to scrutiny that was previously beyond the state's capabilities. While a platform might use a human content moderator to determine whether the uploaded content should be permitted or forbidden, law enforcement uses its investigative tools to determine whether the poster has committed a crime. When law enforcement demands a list of all those users whose uploads were captured by an automated filter, it does not distinguish between them.[253]

The result is that, in complying with this demand, YouTube is providing law enforcement with subscriber information for a fairly broad set of users whose only suspicious act was running afoul of an algorithm. The platforms' efforts to account for context in determining whether a user-generated post is permitted or forbidden are irrelevant to law enforcement.[254] In addition, YouTube's moderation practices make available information about a broader set of actors. The harmful act that once would have prompted law enforcement to monitor or surveil the user's behavior—the actual dissemination of "terrorist content"—has been replaced by the unsuccessful *attempt* to distribute unlawful content.[255] Regardless of the fact that the attempt might, itself, be unlawful as a matter of substantive criminal law, we might question whether these attempts to distribute unlawful content are, by definition, sufficiently grave to warrant law enforcement action.[256]

In a sense, the platforms' ability to cheaply and easily generate a broad array of information relevant to law enforcement is simply an illustration of the extent to which private sector surveillance underpins law enforcement investigations.[257] The platforms' ability—and, indeed, obligation—to generate this information is a classic example of a transformation in technology and society that expands police power by lowering the cost of sur-

[253] *See, e.g.*, *Facebook's Civil Rights Audit*, *supra* note 237, at 10 (noting a variety of pilot efforts to better train human reviewers).

[254] For example, while YouTube tries to differentiate between videos that "discuss topics like pending legislation, aim to condemn or expose hate, or provide analysis of current events," its automated filters are not fully capable of doing so. *See Our Ongoing Work to Tackle Hate*, *supra* note 126.

[255] *Cf.* Heidi R. Gilchrist, *The Vast Gulf Between Attempted Mass Shooting & Attempted Material Support*, 81 U. Pitt. L. Rev. 63, 84–93 (2019).

[256] *See* Shirin Sinnar, *Separate and Unequal: The Law of "Domestic" and "International" Terrorism*, 117 Mich. L. Rev. 1333, 1393 (2019) (noting that many attempted materials to support prosecutions "make[ ] little effort to distinguish between individuals with and without a plausible connection to grave international threats").

[257] *See, e.g.*, Carpenter v. United States, 138 S. Ct. 2206, 2217 (2018) (reasoning that warrantless access to cell site location records in the possession of third-party cell phone companies contravenes societal expectations that police will not "secretly monitor and catalogue every single movement" individuals make).

veillance and expanding the ability to surveil in the first place.[258] These shifts call into question the preexisting balance of power between the government and the public.[259]

But platform surveillance also creates new opportunities for law enforcement precisely because these kinds of mechanisms are not as readily observed—or evaded—as their physical equivalents. To the extent that the new breed of content regulation constrains content moderation practices, it makes platforms a *more* attractive source for law enforcement seeking to obtain information about users. Regardless of whether automated moderation techniques are required by law or simply adopted voluntarily, they increase the wealth of information available about users—their relationships, their interests, and their engagement with online content (all of which platforms already collect)—and that is highly relevant for law enforcement investigations. Moreover, because most of these decisions occur behind closed doors, platform surveillance tends to operate in ways that are not amenable to public oversight or control.[260]

C. Content Moderation as Algorithmic Control

The shift toward automation in content moderation also underscores longstanding concerns about bias, fairness, transparency, and accountability in machine learning and in automated systems more generally.[261] As the private and public sectors increasingly rely on automation in ways that entrench power dynamics and impact individual rights and liberties, these concerns have come to the forefront of scholarship and public discourse.[262]

One concern regards widespread overconfidence in technology itself as a mechanism for solving social and political problems. Magical thinking about artificial intelligence (AI) is prevalent, even though few can even

[258] *See* Orin S. Kerr, *An Equilibrium-Adjustment Theory of the Fourth Amendment*, 125 Harv. L. Rev. 476, 500 (2011).

[259] *Id.*

[260] *Cf.* United States v. Jones, 565 U.S. 400, 415-16 (2012) (Sotomayor, J., concurring) ("[B]ecause GPS monitoring is cheap in comparison to conventional surveillance techniques and, by design, proceeds surreptitiously, it evades the ordinary checks that constrain abusive law enforcement practices: 'limited police resources and community hostility.'") (quoting Illinois v. Lidster, 540 U.S. 419, 426 (2004)).

[261] *See, e.g.*, Jenna Burrell, *How the Machine 'Thinks': Understanding Opacity in Machine Learning Algorithms*, Big Data & Soc'y, Jan.-June 2016, at 1, 4; Paul B. de Laat, *Big Data and Algorithmic Decision-Making: Can Transparency Restore Accountability?*, ACM Computers & Soc'y, Sept. 2017, at 39, 45-46; Sonia K. Katyal, *Private Accountability in the Age of Artificial Intelligence*, 66 UCLA L. Rev. 54, 56 (2019); Joshua A. Kroll et al., *Accountable Algorithms*, 165 U. Pa. L. Rev. 633, 680 (2017); *Fairness, Accountability, and Transparency in Machine Learning*, FAT/ML, https://www.fatml.org [https://perma.cc/249S-ACSC] (last visited Feb. 4, 2020).

[262] *See, e.g.*, Ifeoma Ajunwa et al., *Limitless Worker Surveillance*, 105 Calif. L. Rev. 735, 753 (2017); Solon Barocas & Andrew D. Selbst, *Big Data's Disparate Impact*, 104 Calif. L. Rev. 671, 674 (2016); Ryan Calo & Alex Rosenblat, *The Taking Economy: Uber, Information, and Power*, 117 Colum. L. Rev. 1623, 1627 (2017); Pauline T. Kim, *Data-Driven Discrimination at Work*, 58 Wm. & Mary L. Rev. 857, 863 (2017).

agree on what AI is.[263] Policymakers are not immune to techno-optimism, but have proven vulnerable to its fallacies, embracing innovation by adopting risky technology in settings like healthcare, welfare, education, and criminal justice while failing to regulate AI in any meaningful way.[264] In this regard, AI policy presents a particularly potent example of Edward Felten's Third Law: "lawyers put too much faith in technical solutions, while technologists put too much faith in legal solutions."[265]

But overconfidence in technical solutions can have damaging effects. Far from serving as a neutral arbiter, the algorithms that Internet intermediaries use to rank and prioritize content often reflect and encode social bias.[266] While publicly available displays, such as auto-complete suggestions or search results, can be interpreted as indications of algorithmic bias, other algorithms operate in ways that are more immune from scrutiny.[267] As private platforms determine and control the conditions under which researchers might access the information needed to study algorithmic bias in the first instance, they reinforce their own ability to control political and public narratives regarding algorithmic accountability.[268]

The opacity of content moderation rules, algorithms, and decisions also alters the ways that online users perceive and experience online partic-

[263] *See, e.g.*, Ryan Calo, *Artificial Intelligence Policy: A Primer and Roadmap*, 51 U.C. Davis L. Rev. 399, 406–07 (2017) (describing varying, potential definitions for artificial intelligence); Arvind Narayanan, *How to Recognize AI Snake Oil*, Princeton U. 1, 6 (2019), https://www.cs.princeton.edu/~arvindn/talks/MIT-STS-AI-snakeoil.pdf [https://perma.cc/AME7-TM3P] ("Companies advertising AI as the solution to all problems have been helped along by credulous media.").

[264] *See, e.g.*, Pasquale, *supra* note 22, at 61; Barocas & Selbst, *supra* note 262, at 671; *Technological Due Process*, *supra* note 11, at 1249; Mayson, *supra* note 11, at 2218; Rebecca Wexler, *Life, Liberty, and Trade Secrets: Intellectual Property in the Criminal Justice System*, 70 Stan. L. Rev. 1343, 1346 (2018).

[265] Paul Ohm, *Breaking Felten's Third Law: How Not to Fix the Internet*, 87 Denver U. L. Rev. Online 50, 50 (2010) (quoting Edward Felten).

[266] *See, e.g.*, Safiya Umoja Noble, Algorithms of Oppression: How Search Engines Reinforce Racism 2 (N.Y. Univ. Press 2018); Katyal, *supra* note 261, at 94 (arguing that stereotypes in search results might create a social feedback loop).

[267] For example, some Facebook users have reported a frustrating, Kafkaesque experience of having their accounts disabled for "suspicious activity" without recourse. Kashmir Hill, *Many Are Abandoning Facebook. These People Have the Opposite Problem*, N.Y. Times (Aug. 22, 2019), https://www.nytimes.com/2019/08/22/business/reactivate-facebook-account.html [https://perma.cc/VRC7-MNYD].

[268] Facebook's foundering partnership with Social Science One is a key example of this dynamic. *See, e.g.*, Alex Pasternack, *Frustrated Funders Exit Facebook's Election Transparency Project*, Fast Company (Oct. 28, 2019), https://www.fastcompany.com/90412518/facebooks-plan-for-radical-transparency-was-too-radical [https://perma.cc/E765-TWTX]; Craig Silverman, *Exclusive*: *Funders Have Given Facebook a Deadline to Share Data with Researchers or They're Pulling Out*, BuzzFeed News (Aug. 27, 2019, 4:53 PM), https://www.buzzfeednews.com/article/craigsilverman/funders-are-ready-to-pull-out-of-facebooks-academic-data [https://perma.cc/3HUD-279C]; Gillian Tett, *Why Facebook's Data-Sharing Initiative Is Faltering*, Fin. Times (Nov. 6, 2019), https://www.ft.com/content/98b5385e-0025-11ea-b7bc-f3fa4e77dd47 [https://perma.cc/CA9X-YU36].

ipation.[269] On the one hand, the apparent arbitrariness of moderation decisions—and of algorithmic moderation in particular—might heighten the need for transparency in order to promote legitimacy. In response to shadowy decisions about user-generated content, users have developed what Sarah Myers West has charitably called "folk theories" regarding moderation.[270] More transparency might promote user trust in the system, but it depends on how meaningful the disclosures are.[271] On the other hand, meaningful transparency for *ex ante*, automated moderation techniques would also undermine their effectiveness, likely resulting in more *ex post* screening. If transparency enabled users to "reverse engineer" moderation standards, it may result in users developing a novel, new vocabulary to evade moderation—entrenching unwanted content while making it more difficult to detect.[272]

Policymakers' faith in the power of private innovation is perhaps most visible in contexts such as statutes that require technology companies to invest heavily in new, untested technologies of moderation, or court rulings that assume platforms have technical capabilities not in evidence.[273] But when it comes to content moderation, much of this faith is misplaced. For example, in the aftermath of the Christchurch shooting, YouTube and Facebook users altered the footage slightly—for example, by surrounding it in a frame, or by posting video of the footage streaming in a second window—in order to get the footage past automated detection mechanisms.[274] These incidents illustrated how automated content moderation systems struggle to draw lines between protected and illicit content.

These failures prompted widespread outrage, but addressing the problem using mechanical solutions would require greater efforts to suppress users' posts. Moreover, broader approaches to preventing the dissemination of unlawful content might result in platforms suppressing newsworthy posts in addition to "gratuitous graphic violence."[275] For content that inherently lacks any redeeming social value, such as child pornography or non-consensual pornography, the need to prevent harm might justify the

[269] *See* J. Nathan Matias, *The Civic Labor of Volunteer Moderators Online*, Soc. Media & Soc'y, Apr.-June 2019, at 1, 4-5; Nicolas P. Suzor et al., *What Do We Mean When We Talk About Transparency? Toward Meaningful Transparency in Commercial Content Moderation*, 13 Int'l J. Comm. 1526, 1527 (2019) [hereinafter *What Do We Mean When We Talk About Transparency?*]; *Censored, Suspended, Shadowbanned*, *supra* note 52, at 4366; Sarah Myers West, *Raging Against the Machine: Network Gatekeeping and Collective Action on Social Media Platforms*, Media & Comm., Sept. 2017, at 28, 28.

[270] *Censored, Suspended, Shadowbanned*, *supra* note 52, at 4372-74.

[271] Suzor et al., *supra* note 269, at 1527.

[272] *See* Chancellor et al., *supra* note 98, at 1201 (explaining how pro-eating disorder communities on Instagram adopted unusual spellings to circumvent hashtag moderation techniques).

[273] *See supra* Part II.

[274] Issie Lapowsky, *Why Tech Didn't Stop the New Zealand Attack from Going Viral*, Wired (Mar. 15, 2019, 1:50 PM), https://www.wired.com/story/new-zealand-shooting-video-social-media [https://perma.cc/VE3L-MMV5].

[275] *Id.*

cost to free expression.[276] But in other contexts, such as defamation, hate speech, or terrorist propaganda, whether content is unlawful "depends on the overall context, including the message and precise wording."[277]

The challenge of designing automated systems to identify and suppress certain kinds of content highlights the more straightforward political difficulty of defining unlawful content in the first instance. Even if platforms could automate the detection of terrorist propaganda, extremist content, or hate speech, defining those categories will be as difficult as identifying parody, fake news, or fair use.[278] This inquiry is fact-bound and culturally specific. With respect to all but the clearest cases, policymakers and platforms will find it difficult to apply these distinctions and determine the social value of user-generated content. That this is a fundamental problem of free expression explains why automated decision-making alone cannot answer the challenge.

Of course, forcing humans to decide whether horrific content ought to be permitted or taken down creates its own problems. Content moderators often work in "sweatshop-like" conditions to clean up the Internet.[279] Content moderators often experience serious trauma from viewing so many disturbing posts and images in quick succession.[280] Some of them come to believe conspiracy theories expressed in moderated content.[281] And contractors sometimes discourage moderators from raising questions to Facebook about unclear subjects, perpetuating the lack of clarity and apparent arbitrariness of some of these rules.[282]

The human cost of content moderation will amplify calls for automated moderation techniques. Facebook and other platforms play into this narrative, arguing that artificial intelligence can help platforms solve their content-related problems.[283] But the platform may not be able to avoid using human content moderators. For example, under the General Data Protection Regulation (GDPR), individuals "have the right not to be subject to a decision based *solely* on automated processing . . . which produces legal effects concerning him or her or similarly significantly affects

[276] Danielle Citron & Quinta Jurecic, *Platform Justice* 11–12 (Hoover Inst. Aegis Series Paper No. 1811, 2018).

[277] *Id.* at 13.

[278] *The Platform Is the Message*, *supra* note 31, at 15 ("[I]dentifying and excluding fake news is a hard line-drawing problem.").

[279] *See* Hanna Kozlowska, *This Documentary Shows the Sweatshop-Like Labor of Internet Content Moderators*, QUARTZ (Nov. 12, 2018), https://qz.com/1460906/the-cleaners-is-a-documentary-that-shows-the-sweatshop-like-labor-of-content-moderators/ [https://perma.cc/XC56-AZ6F].

[280] Casey Newton, *The Trauma Floor: The Secret Lives of Facebook Moderators in America*, VERGE (Feb. 25, 2019, 8:00 AM), https://www.theverge.com/2019/2/25/18229714/cognizant-facebook-content-moderator-interviews-trauma-working-conditions-arizona [https://perma.cc/W6K3-ZD3J].

[281] *Id.*

[282] *Id.*

[283] James Vincent, *AI Won't Relieve the Misery of Facebook's Human Moderators*, VERGE (Feb. 27, 2019, 12:41 PM), https://www.theverge.com/2019/2/27/18242724/facebook-moderation-ai-artificial-intelligence-platforms [https://perma.cc/6EXK-D4H8].

him or her."[284] Even those who are most optimistic about the use of artificial intelligence to detect harmful online content acknowledge the vital role that human moderators play in deciding whether user-generated posts ought to remain online or be taken down.[285]

D. Content Moderation as Power

Finally, automation mandates may entrench the position of firms at the leading edge of AI development, such as Facebook and Google. In a climate of increasing sensitivity to technology platforms' market dominance, content regulation has proven attractive for opponents of "Big Tech."[286] However, some have argued that automation requirements are as burdensome as to exclude new start-ups and smaller competitors.[287] Moreover, as the developers of AI moderation tools, technology platforms such as Google, Facebook, and Twitter, and companies such as Crisp, Twohat, and Adobe stand to gain substantially from the expansion of automation mandates.[288]

A full assessment of how content regulation might reshape competition is beyond the scope of this Article. But the experience of algorithmic, copyright enforcement confirms that commercial interests provide powerful motives for platforms to monetize their moderation systems by selling their services to competitors.[289] As Bridy documented, Audible Magic—one of the chief purveyors of copyright filtering technology—lobbied the European Commission to require proactive filtering in Article 17.[290]

[284] Council Regulation 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC, art. 22, 2016 O.J. (L 119) 15, 65 (EU) [hereinafter GDPR].

[285] Telephone Interview with Mark Zuckerberg, CEO, Facebook, Inc. (Apr. 2, 2019), *available at* https://www.fool.com/earnings/call-transcripts/2019/04/25/facebook-inc-fb-q1-2019-earnings-call-transcript.aspx [https://perma.cc/ZR42-YYTT] ("The only hope is building AI systems that can either identify things and handle them proactively or at the very least, flag them for a lot of people who work for us who can then look at them.").

[286] Tim Mak, *Senator Pushes Bill to Curb 'Exploitative and Addictive' Social Media Practices*, NPR (Aug. 14, 2019, 5:00 AM), https://www.npr.org/2019/08/14/750585438/senator-pushes-bill-to-curb-exploitative-and-addictive-social-media-practices [https://perma.cc/S8HQ-E8AG].

[287] *See* James Temperton & Matt Reynolds, *The European Parliament Has Voted in Favour of Article 13*, WIRED (Mar. 26, 2019), https://www.wired.co.uk/article/eu-article-13-vote-article-17 [https://perma.cc/8MER-6MSF].

[288] One major content moderation contractor, Cognizant, announced in fall 2019 that it would exit the content moderation business, while allocating $5 million to research on automation and algorithmic moderation. Telephone Interview with Brian Humphries, CEO, Cognizant Technology Solutions Corp. (Oct. 30, 2019), *available at* https://www.fool.com/earnings/call-transcripts/2019/10/31/cognizant-technology-solutions-corp-ctsh-q3-2019-e.aspx [https://perma.cc/M6S6-9LE3].

[289] *See, e.g.*, BFDI Press Release, *supra* note 182 (raising concerns that large platforms will sell their moderation systems to small platforms, cementing their dominance).

[290] *See* Annemarie Bridy (@AnnemarieBridy), TWITTER (Jan. 19, 2019, 6:06 PM), https://twitter.com/AnnemarieBridy/status/1086761804301094917 (citing a promo-

Already, companies such as Adobe[291] and Crisp[292] are selling proprietary filtering software that purports to effectively filter hate speech and terrorist content. Proactive monitoring requirements will create a vast, new market for automated moderation techniques. These market effects will be present regardless of whether new regulations *require* the use of *ex ante* screening methodologies, or simply encourage them.

E. Content Moderation as Extraterritorial Governance

Pushing platforms to adopt automated, *ex ante* screening methodologies is a matter of global significance. These changes send a signal to other governments—and to lobbyists—that domestic law can effectively require online service providers to be more proactive in filtering and monitoring content.[293] As platforms build and acquire the technological infrastructures necessary to comply with European law, they are likely to use those infrastructures on a global scale, not only in their European affiliates.[294]

Local content regulations have global effects in part because platforms prefer to enforce their terms of service before enforcing local law.[295] As platforms increasingly automate the screening of user-generated content for compliance with their terms of service, content that violates community guidelines will be taken down globally.[296] The global effects of takedowns combine with the problems of over-deletion. For instance, when service providers receive DMCA notices, they generally delete the content across the entire platform, rather than by blocking or filtering content within the U.S.[297]

tional video regarding "content recognition technologies" in Article 17—previously referred to as Article 13).

[291] *See, e.g.*, *Content Management Meets AI with Adobe Experience Manager Sites*, Adobe, https://www.adobe.com/marketing-cloud/experience-manager/social-media-moderation.html [https://perma.cc/SX4T-77US] (last visited Apr. 16, 2020).

[292] *See, e.g.*, *Crisp Solutions: Digital Marketing Defense*, Crisp Thinking, https://www.crispthinking.com/content-moderation-for-social-networks [https://perma.cc/DY3G-GR4D] (last visited Apr. 16, 2020).

[293] *See, e.g.*, Anu Bradford, *The Brussels Effect*, 107 Nw. U. L. Rev. 1, 44 (2012) (citing Katerina Linos's work on legal diffusion). Indeed, organizations that represent rightsholders have already urged the United States Copyright Office to consider filtering and fingerprinting technologies as "standard technical measures" that should be required in order for companies to benefit from Section 512's safe harbor. *See also, e.g.*, Copyright Office of the United States of America, Comment Letter on Section 512 Study: Notice and Request for Public Comment, submitted by Jay Rosenthal & Steven Metalitz (Apr. 1, 2016), http://static.politico.com/a3/bf/686b5f2942dbb2b5327a8a2d8ceb/music-community-submission-in-re-dmca-512-2.pdf [https://perma.cc/FXZ9-NDWV].

[294] *See* Bloch-Wehba, *supra* note 21, at 29 (describing how platform terms of service generally apply globally, not locally).

[295] *See, e.g.*, Amélie Heldt, *Reading Between the Lines and the Numbers: An Analysis of the First NetzDG Reports*, Internet Pol'y Rev., June 2019, at 1, 11 (documenting how platforms enforcing Germany's NetzDG law tend to screen content against their community guidelines before assessing compliance with local law).

[296] *See also Extremist Speech*, *supra* note 10, at 1055–57.

[297] Annemarie Bridy, *Intellectual Property*, *in* Law, Borders, and Speech: Proceedings and Materials 13 (Daphne Keller ed., Stanford Ctr. for Internet & Soc'y 2017), *available at* https://cyberlaw.stanford.edu/files/publication/files/12-18%20FINAL%20

These practices raise potent questions about the implications of new content regulations for jurisdictions. When governments misappropriate the instruments of private governance, they can often achieve—whether intentionally or not—global effects for public policy. While issues such as data localization, data sovereignty, and extraterritoriality remain in the public eye, it is worth considering how content regulation itself manifests many of the same controversial aspects, like calling into question the consent of the governed.[298]

IV. Principles for Moderation of Automation

By design, the new automated technologies of moderation not only promote and advance the interests of law enforcement, but they also cultivate a sense of mutual dependency in which states require platforms' assistance to vindicate policy objectives and platforms comply in order to avoid harder regulation. Although this arrangement might seem "efficient," in the sense that both, states and platforms, receive optimal outcomes without costly interventions, they pose a threat to democratic values and safeguards. While artificial intelligence and other automated content moderation tools hold substantial promise for scaling the work of content moderation, they come at a significant cost to civil liberties and are poised to entrench the power of the private sector.

This Part concludes by pointing toward several ways in which regulation might seek to exert a moderating influence upon the use of automation itself. By recalibrating the regulatory balance away from the current emphasis on "scalability," legislation and regulation might reach a healthier resolution that squares the challenges of harmful online content with public governance and with individual rights.

A. Platform Transparency

Several of the new measures require platforms to produce annual or semi-annual transparency reports documenting the actions they have taken to address unlawful content.[299] But while transparency reports originated as a way for technology companies to notify *users* about demands for surveillance and censorship, these new reporting obligations operate as a way for *governments* to ensure that platforms are keeping up with the uptick in

Conference%20Proceedings.pdf [https://perma.cc/F8VJ-WFJ9] ("Has notice and takedown for copyright become de facto harmonized through platforms' global application of the DMCA?").

[298] Bloch-Wehba, *supra* note 21, at 67-68 (describing questions about legitimacy that arose from the emergence of governance institutions outside the state).

[299] *See, e.g.*, NETZWERKDURCHSET ZUNGSGESETZ [NetzDG] [Network Enforcement Act], Oct. 1, 2017, BUNDESMINISTERIUM DER JUSTIZ UND FÜR VERBRAUCHERSCHUTZ [BMJV] at art. 1(2) (Ger.) (requiring social networks that receive over 100 complaints per year to produce semi-annual transparency reports); *Draft Terrorism Regulation*, *supra* note 189, at art. 8(2) (requiring service providers that have received removal orders to publish annual transparency reports).

censorship demands.[300] In other words, transparency reports once served companies' public relations goals by establishing their independence from the government. Now, in an era of concern about platform impunity, companies release data about content moderation in order to show compliance with government demands.[301]

The existing transparency protections are woefully incomplete. The transparency-reporting obligations under NetzDG and the draft terrorism regulation require platforms to document their efforts to take down content that is the subject of a specific complaint, but do not fully capture companies' decisions to deploy automated mechanisms that avoid complaints arising in the first place. For this reason, platforms ought to be more transparent about how, when, and why they deploy *ex ante*, automated screening of user-generated content. Precisely because laws like NetzDG, the Australian AVM law, and Article 17 of the Copyright Directive encourage—but do not require—automated content recognition, platforms have choices about whether or not to do so. Understanding how automated measures fit within the framework of content moderation is necessary to have a fuller picture of the relationship between government pressure and private sector practices.[302]

Perhaps more importantly, as Daphne Keller has pointed out, platform transparency reports can only present a highly limited perspective on content moderation that reflects "the platforms' own characterization of the content they took down."[303] In its current form, aggregate data does not explain, for example, the kinds of content that are swept up by an algorithm designed to detect "terrorism," nor the reasons that a platform might not identify "white nationalism" as "terrorist content."[304]

More robust transparency practices might shed some much-needed light on these shifting dynamics. For example, platforms might routinely

[300] *See* Jonathan Manes, *Online Service Providers and Surveillance Law Transparency*, 125 YALE L.J. F. 343, 344 (2016), https://www.yalelawjournal.org/forum/online-service-providers-and-surveillance-law-transparency [https://perma.cc/DE3D-38HL] ("If these companies could win the right to speak about the *kinds* of records the government is ordering them to disclose, they would be able to provide the public with crucial information about how the surveillance laws have been interpreted and applied in practice.").

[301] Mike Masnick, *How Government Pressure Has Turned Transparency Reports from Free Speech Celebrations to Censorship Celebrations*, TECHDIRT (Apr. 17, 2018, 12:04 PM), https://www.techdirt.com/articles/20180402/07014939543/how-government-pressure-has-turned-transparency-reports-free-speech-celebrations-to-censorship-celebrations.shtml [https://perma.cc/85EX-V782].

[302] To its credit, the draft terrorism regulation requires government agencies as well as platforms to record removal orders. *See Draft Terrorism Regulation*, *supra* note 189, at art. 8(a). Unfortunately, this requirement is skeletal, calling for governments to disclose data regarding removal orders, investigations, and content "wrongly identified as terrorist." *Id.* In addition, government authorities must disclose a "description of measures" requested from service providers. *Id.*

[303] *See Three Constitutional Thickets*, *supra* note 185, at 8.

[304] Díaz, *supra* note 129 (expressing the concern that GIFCT's definition of "glorification" of terrorism is "imprecise"); Nitasha Tiku, *Tech Platforms Treat White Nationalism Different from Islamic Terrorism*, WIRED (Mar. 20, 2019, 8:00 AM), https://www.wired.com/story/why-tech-platforms-dont-treat-all-terrorism-same/ [https://perma.cc/SB88-WL7W].

review and audit their algorithms and datasets to determine whether automated methods experience different error rates with different speakers, languages, or contexts, and then, publish the results.[305] By the same token, platforms might disclose other statistical information that would show, for example, whether some types of speech garnered disproportionate complaints under NetzDG or other statutory mechanisms.

However, even robust audits and voluntary disclosures will lose credibility if they are self-enforced and self-policed. Therefore, in addition to requirements that platforms publish aggregate data, regulators might consider requiring *algorithmic* transparency mechanisms.[306] A growing body of work has begun to critique the power of technology firms to "lock away information in the face of a strong public interest in disclosure."[307] Without legislative intervention, platforms are likely to treat their methodologies of algorithmic enforcement as "trade secrets," just as they have vigorously sought to shield their policies on content moderation from public disclosure.[308]

A full review of the extensive literature on algorithmic transparency and accountability is beyond the scope of this Article.[309] For our purposes, it suffices to say that a regulator might opt to include provisions that would make automated, *ex ante* content screening less inscrutable, either by providing for government audits, facilitating independent research, or by requiring disclosure.[310]

Transparency alone is not enough to ensure accountability, of

[305] Indeed, major technology platforms are at the forefront of research on artificial intelligence and machine learning, and, given the current distributions of resources, these platforms are likely the institutions best equipped to undertake these kinds of projects. *See, e.g.*, AMAZON, https://www.aboutamazon.com/research [https://perma.cc/CUA3-SKCB] (last visited Apr. 16, 2020); GOOGLE, https://ai.google/research/ [https://perma.cc/9ZLE-ZU2W] (last visited Apr. 16, 2020); FACEBOOK, https://research.fb.com/ [https://perma.cc/9ELJ-ANFQ] (last visited Apr. 16, 2020); MICROSOFT, https://www.microsoft.com/en-us/research/ [https://perma.cc/L2RK-YAJ9] (last visited Apr. 16, 2020).

[306] *See, e.g.*, Mike Ananny & Kate Crawford, *Seeing Without Knowing: Limitations of the Transparency Ideal and Its Application to Algorithmic Accountability*, 20 NEW MEDIA & SOC'Y 973, 974 (2016); *Technological Due Process*, *supra* note 11, at 1260; *Transparency and Algorithmic Governance*, *supra* note 11, at 4.

[307] Oren Bracha & Frank Pasquale, *Federal Search Commission—Access, Fairness, and Accountability in the Law of Search*, 93 CORNELL L. REV. 1149, 1202 (2008). *See also* Hannah Bloch-Wehba, *Access to Algorithms*, 88 FORDHAM L. REV. 1265, 1290 (2020); Dan L. Burk & Julie E. Cohen, *Fair Use Infrastructure for Rights Management Systems*, 15 HARV. J.L. & TECH. 41, 59 (2001).

[308] Burk & Cohen, *supra* note 307, at 67.

[309] *See, e.g.*, Burrell, *supra* note 261, at 10; Robert Brauneis & Ellen P. Goodman, *Algorithmic Transparency for the Smart City*, 20 YALE J.L. & TECH. 103, 110 (2018); John Danaher, *The Threat of Algocracy: Reality, Resistance and Accommodation*, 29 PHIL. & TECH. 245, 258 (2016); Nicholas Diakopoulos, *Accountability in Algorithmic Decision Making*, COMM. ACM, Feb. 2016, at 56, 60.

[310] *See, e.g.*, *What Do We Mean When We Talk About Transparency?*, *supra* note 270, at 1529 (calling for more disclosure of disaggregated data with independent researchers). *See also* Rory Van Loo, *Regulatory Monitors: Policing Firms in the Compliance Era*, 119 COLUM. L. REV. 369, 424 (2019).

course.[311] That is why it is critical to pair transparency obligations with other commitments to public oversight and to check the power of the private sector, including ongoing monitoring.[312] But those who seek to censor and surveil also benefit from *sub rosa* arrangements that blur the line between the private and public sectors. Shedding light on those arrangements is integral to holding these powerful institutions accountable.

B. Procedural Safeguards

Platforms and their regulators might also consider embracing more robust procedural safeguards that protect users who contest blocking and filtering decisions. As Balkin recognized, digital filtering systems might tend to operate as "prior restraints" on speech that prevent individuals from speaking, rather than punishing them after-the-fact.[313] Just as prior restraints call for specific kinds of procedural protections to guard against the risk of censorship, regulators might likewise integrate procedural safeguards—such as appeal mechanisms and judicial review requirements—into platform regulation.[314]

Appeal mechanisms have gained substantial traction, especially as Facebook began ramping up its Oversight Board to review its content moderation practices.[315] These mechanisms need not be strictly private or voluntary. For example, the Copyright Directive requires each platform to create an appeal mechanism for users to contest the removal of their content.[316] Similarly, under the GDPR, individuals "have the right not to be subject to a decision based solely on automated processing."[317]

[311] For critiques of this fallacy, see, e.g., Margaret B. Kwoka, *FOIA, Inc.*, 65 Duke L.J. 1361, 1365 (2016) (documenting how frequently the private sector uses transparency law in its own self-interest); David E. Pozen, *Freedom of Information Beyond the Freedom of Information Act*, 165 U. Pa. L. Rev. 1097, 1146 (2017); Andrew Keane Woods, *The Transparency Tax*, 71 Vand. L. Rev. 1, 3 (2018). *See also* Woodrow Hartzog, *Body Cameras and the Path to Redeem Privacy Law*, 96 N.C. L. Rev. 1257, 1275-77 (2018) (describing, in the context of body cameras, the concern that public transparency might be coopted for private sector gain).

[312] *What Do We Mean When We Talk About Transparency?*, *supra* note 269, at 1527-28; Rory Van Loo, *The Missing Regulatory State: Monitoring Businesses in an Age of Surveillance*, 72 Vand. L. Rev. 1563, 1595 (2019) (suggesting that the ongoing monitoring of platforms' speech practices might be justified).

[313] Jack Balkin drew an analogy between digital filtering systems and prior restraints. Relying on this comparison, certain safeguards might be appropriate in the former context precisely because of their similarity to the latter. *See Old-School/New-School Speech Regulation*, *supra* note 12, at 2318.

[314] *Id. See also* Freedman v. Maryland, 380 U.S. 51, 58-59 (1965) (articulating three safeguards).

[315] Casey Newton, *Facebook's Oversight Board Could Bring a Justice System to a Platform That Needs One*, Verge (Sept. 18, 2019, 6:00 AM), https://www.theverge.com/interface/2019/9/18/20870605/faceboook-oversight-board-charter-justice-system [https://perma.cc/B37U-V4WQ].

[316] Copyright Directive 2019/790, *supra* note 165, at art. 17(9).

[317] GDPR, *supra* note 284, at art. 22(1). *See also* Emily Pehrsson, *The Meaning of the GDPR Article* 22 1, 22 (Stanford-Vienna Transatlantic Tech. Law Forum, Working Paper No. 31, 2018) (discussing whether Article 22 is an outright prohibition of automated decision-making or confers a "right to challenge" the outcome of an automatic decision).

Apart from substantive oversight of content-related decisions, appeals are also important because they play an integral role in promoting transparency and legitimacy. Through an appeals process, platforms might disclose information about the reasons that a piece of content is blocked or taken down to the individual users who are affected. Thus, through appeals, platforms (at least in theory) engage in a familiar kind of administrative reason-giving.[318]

Appeal mechanisms have several major drawbacks, however. First, while they might make marginal improvements to transparency, they are opaque and ineffective protections against over-deletion.[319] Second, like moderation itself, appeals present a problem of scale. Finally, requiring platforms to create expensive and burdensome appeal mechanisms threatens small companies and start-ups while favoring dominant incumbents.

Large online platforms can and do construct entire quasi-legal regimes for online speech, replete not only with statutes and regulations (terms of service and community guidelines), but also with legal structures (complaints and appeals). Facebook's Oversight Board (the Board) illustrates how one large company has approached this issue by designing an independent body to oversee its content moderation decisions.[320] In November 2018, Mark Zuckerberg announced that Facebook would create "a new way to appeal content decisions to an independent body."[321] The company then opened a "public consultation process" for six weeks to get public feedback on the design of the Board.[322] After holding a series of invitation-only workshops and roundtables, the platform published its draft Charter in September 2019.[323] Under the Charter, the Board has the authority to consider appeals of content-related decisions.[324] The Board can also set its own mechanisms for determining which "cases" to

[318] Frank I. Michelman, *Formal and Associational Aims in Procedural Due Process*, 18 NOMOS 126, 126 (1977). As Michelman points out, even what he calls "nonformal" explanations have significance for due process: they "seem responsive to demands for *revelation* and *participation*. They attach value to the individual[] *being told why* the agent is treating him unfavorably and to his *having a part in the decision*." *Id.* at 127.

[319] *See, e.g.*, *What Do We Mean When We Talk About Transparency?*, *supra* note 269, at 1537; URBAN ET AL., *supra* note 228, at 58; *Censored, Suspended, Shadowbanned*, *supra* note 52, at 4378–79.

[320] Kate Klonick & Thomas Kadri, *How to Make Facebook's 'Supreme Court' Work*, N.Y. TIMES (Nov. 17, 2018), https://www.nytimes.com/2018/11/17/opinion/facebook-supreme-court-speech.html [https://perma.cc/MYW9-QRX4].

[321] Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement,* FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerburg/a-blueprint-for-content-governance-and-enforcement/10156443129621634 [https://perma.cc/773G-CYQN].

[322] Brent Harris, *Getting Input on an Oversight Board*, FACEBOOK (Apr. 1, 2019), https://about.fb.com/news/2019/04/input-on-an-oversight-board/ [https://perma.cc/LXG2-DNRK].

[323] Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, FACEBOOK (Sept. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure/ [https://perma.cc/3GJL-T55E].

[324] *Oversight Board Charter*, FACEBOOK 1, 4–5 (2019), https://fbnewsroomus.files.wordpress.com/2019/09/oversight_board_charter.pdf [https://perma.cc/JCD9-Q8YG].

consider.[325]

Despite the lengthy process for constructing the Board, the Charter is strikingly short on detail regarding some essential aspects of its procedures. For example, the Charter is silent on whether parties before the Board can be represented by counsel.[326] The narrow scope of the Board's jurisdiction—only individual content decisions can be appealed—is also questionable. For instance, the Charter does not include specific provisions regarding user appeals from decisions to disable their accounts, even though that might also be considered a content-related decision.[327] Since Facebook does not offer appeals for all content-related decisions, there are presumably also some areas over which the Board will lack authority, such as child sexual abuse imagery.[328]

The Charter also anticipates that the Board's decision in one appeal might be binding on other content as well. First, the Board's decisions are "precedential."[329] Yet, the Charter also notes that, "where Facebook identifies that identical content with parallel context—which the Board has already decided upon—remains on Facebook, it will take action by analyzing whether it is technically and operationally feasible to apply the Board's decision to that content as well."[330] Thus, while the Board will lack authority to decide cases arising from Facebook's algorithmic delivery, curation, or ranking of content, the Charter also anticipates that the company might use its technical tools to instantiate Board decisions.[331]

The effectiveness of Facebook's new appeals mechanisms largely depends on factors that have yet to be publicly announced—namely, whom it appoints to the Board.[332] And Facebook's commitment to public participation, input, and careful drafting in the process of formulating the Charter does not ensure that the Board's approach to content governance will add anything more than a symbolic veneer of compliance with free expres-

[325] *Id.* at 5.

[326] *See generally id.*

[327] Sarah C. Haan, *Bad Actors: Authenticity, Inauthenticity, Speech, and Capitalism*, U. Pa. J. Const. L. 1, 47 (forthcoming) ("[C]ompanies sometimes justify this approach on the ground that 'authentic' speakers produce 'authentic content,' which implies that content produced by authentic speakers is truthful and good.").

[328] *Understanding the Community Standards Enforcement Report*, Facebook, https://transparency.facebook.com/community-standards-enforcement/guide [https://perma.cc/D6NY-S4WV] (last visited June 22, 2020) ("[W]e offer appeals for the vast majority of violation types on Facebook. We don't offer appeals for violations with extreme safety concerns . . . .").

[329] *Oversight Board Charter*, *supra* note 324, at 5.

[330] *Id.* at 7.

[331] Dipayan Ghosh, *Facebook's Oversight Board Is Not Enough*, Harv. Bus. Rev. (Oct. 16, 2019), https://hbr.org/2019/10/facebooks-oversight-board-is-not-enough [https://perma.cc/X5PR-4H4R] (calling for "oversight of the company's algorithmic decision-making to protect against bias").

[332] Jonathan Zittrain, *A Jury of Random People Can Do Wonders for Facebook*, Atlantic (Nov. 14, 2019), https://www.theatlantic.com/ideas/archive/2019/11/let-juries-review-facebook-ads/601996/ [https://perma.cc/UP5R-2PVY] ("A bunch of retired judges or other thoughtful people on that board can, perhaps, deliberate, show their reasoning, and thus convince even those who don't agree with them that the process wasn't rigged against them.").

sion values.[333] Facebook's effort to craft a participatory process—and its invocation of the analogy to a "Supreme Court"[334]—does not change the fact that this is a simulacrum of due process, unregulated by law or the Constitution, and therefore, unaccountable to the democratic process. In fact, Facebook's grand experiment in constitutionalism just highlights that platforms are free to design their quasi-legal protections without any legal consequences or guarantees.

Partly in response to these concerns, platforms, non-governmental organizations, and other stakeholders have considered a range of alternative options for private regulation to help rectify the imbalance, including multi-stakeholder Social Media Councils, or SMCs.[335] SMCs are similar to Facebook's Oversight Board in the sense that they are soft-law institutions, and that their success relies upon voluntary adherence.[336] But in other respects, the similarity runs out. Rather than the Facebook Oversight Board's adjudicatory model, which focuses primarily on hearing individual user appeals, SMC proposals have focused on a multi-stakeholder model that would represent civil society organizations, platforms, users, and governments and advise them on content moderation issues far beyond takedowns.[337] Rather than being led by a single platform, SMCs can offer guidance on cross-cutting issues affecting multiple platforms or the social media sector more generally.[338] SMCs might be global, national, or regional in scope.[339]

Perhaps the most significant aspect of the SMC proposals is that, unlike the Facebook Oversight Board, the fundamental business model of the platform need not be off-limits to the SMC. One can imagine a world in which the SMC's adjudicatory and advisory functions go well beyond what platforms define as "content moderation" and address other issues as well: algorithmic ranking, advertising policy, and anonymity—to name just a few. The potential breadth of the SMC concept, in turn, highlights how narrow Facebook's mandate for the Oversight Board truly is. Precisely because SMCs are envisioned as multi-stakeholder institutions, they may have greater potential to shed light on the entwined relationships between

[333] Carolynn M. Ryan, Internal Dispute Resolution 2 (IRC Press 1998), *available at* https://irc.queensu.ca/sites/default/files/articles/communicating-during-an-organizational-change.pdf [https://perma.cc/ECJ4-N7XJ] (observing that internal dispute resolution entails greater corporate control).

[334] Klonick & Kadri, *supra* note 320.

[335] Land, *supra* note 10, at 57-59 (documenting different accountability mechanisms).

[336] Eileen Donahoe et al., *Social Media Councils: From Concept to Reality*, Stan. GDPi 1, 24 (2019), https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/gdpiart_19_smc_conference_report_wip_2019-05-12_final_1.pdf [https://perma.cc/32QQ-J9Y6] (describing how discussion of appointments to SMCs was partly a response to Facebook's announcement that it would appoint the first Oversight Board).

[337] *Id.* at 26-32 (presenting two potential models for SMCs). While the Article 19 model emphasizes an adjudicatory role for SMCs, the GDPi model emphasizes their advisory function. *Id.*

[338] *Id.*

[339] *Id.*

platforms and states. At the same time, however, SMCs would likely find it difficult to maintain financial independence without significant state and private-sector sponsorship, calling their neutrality into question.

These proposals warrant fuller consideration, especially regarding their potential to realign—or reaffirm—the relationship between user speech, private power, and government censorship. In particular, the emergence of powerful corporate and multi-stakeholder institutions for resolving speech issues might raise questions about the long-standing assumption that government intervention is more dangerous to free speech than private action.[340] But, at least for now, the emergence of the Facebook Oversight Board has dominated discussion of the potential for soft-law institutions to intervene in content moderation debates.[341] It remains to be seen whether these new governance structures create more or less powerful safeguards against wrongful deletion, censorship, and surveillance.

1. *Court Orders*

In addition to appeal mechanisms, as Molly Land noted, many proposals to improve accountability for content moderation have focused on the need for formal legal processes—subject to judicial review—before a state can request that a platform delete content.[342] Without formal mechanisms for *ex ante* judicial review and *ex post* remedies, government demands pose the serious risk of coopting not only platforms' substantive decisions, but also their rules, regulations, and internal decision-making procedures.

So-called IRUs, the law enforcement squadrons that flag illicit content online under private terms of service, highlight these risks. Most online platforms require a court order or other formal request to justify complying with a law enforcement demand to remove user content.[343] However, IRUs operate as if they were ordinary users, flagging violations of the community standards just like any other individual.[344] By employing a plat-

[340] *See, e.g.*, Jerome A. Barron, *Access to the Press—A New First Amendment Right*, 80 Harv. L. Rev. 1641, 1642 (1967) ("[O]nly by responding to the present reality of the mass media's repression of ideas can the constitutional guarantee of free speech best serve its original purposes.").

[341] *See generally, e.g.*, Evelyn Douek, *Facebook's Oversight Board: Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & Tech. 1 (2019); Kate Klonick, *The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 Yale L.J. 2232 (2020); Catalina Botero-Marino et al., *We Are a New Board Overseeing Facebook. Here's What We'll Decide*, N.Y. Times (May 6, 2020), https://www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html [https://perma.cc/TU8U-62PN]; Newton, *supra* note 315.

[342] Land, *supra* note 10, at 57.

[343] *See, e.g.*, *Google Transparency Report*, Google, https://transparencyreport.google.com/government-removals/overview?hl=en [https://perma.cc/HAE6-UHG2] (last visited July 7, 2020) (describing types of requests from government entities); *but see Facebook Transparency Report*, Facebook, https://transparency.facebook.com/content-restrictions [https://perma.cc/6ENF-P4TA] (describing how government entities sometimes request the deletion or restriction of unlawful content without specifying which method to use).

[344] Land, *supra* note 10, at 23–24.

form's community standards rather than the law itself, IRUs can achieve an end-run around legal constraints meant to guard against censorship: they can avoid judicial review and oversight. Obviously, this easier and more straightforward strategy is enticing, but its costs to free expression are significant.

In response, advocates and scholars have stressed the urgency of requiring court orders and formal legal processes before state actors can demand content takedowns.[345] While this approach is an important constraint on discrete government demands that might relate to individual users, posts, and pages, I am not confident that it addresses the broader dynamics raised by automation in moderation. In particular, discrete court orders are unlikely to address the greater tendency of platforms to adopt technological solutions and symbolic structures of compliance to avoid harder regulation. Additionally, maintaining a focus on discrete government demands risks overlooking the emerging pressures on platforms to create new technologies and techniques of moderation. These new forms of government pressure might not take the same shape as the old demands to censor or surveil, but they will affect platforms' design choices and their implementation of private governance structures.

In other words, the risk that government actors might use informal or coercive processes to restrict speech and privacy is not limited to content takedowns or user-information demands but have increasingly extended to the design and implementation of platform rules and compliance systems. These kinds of coercive maneuvers are particularly powerful because they are part of a constellation of simultaneous, increasing pressures on platforms.[346] As a result, judicial orders for takedowns, while an important constraint on the state's ability to demand that platforms carry out its censorship and surveillance objectives, seem ill-equipped to address the risk that platforms might overcompensate in order to seem eager to comply.[347]

[345] *Id.* at 64. *See also* Chang, *supra* note 132, at 124–25.

[346] For example, anti-encryption measures have been enacted in the United Kingdom and Australia. *See* Lily Hay Newman, *Australia's Encryption-Busting Law Could Impact Global Privacy*, Wired (Dec. 7, 2018, 12:45 PM), https://www.wired.com/story/australia-encryption-law-global-impact/ [https://perma.cc/Z49X-LKZ2]. Moreover, states are increasingly demanding that platforms store data locally. *See, e.g.*, Ronak D. Desai, *India's Data Localization Remains a Key Challenge for Foreign Companies*, Forbes (Apr. 30, 2019, 2:35 AM), https://www.forbes.com/sites/ronakdesai/2019/04/30/indias-data-localization-remains-a-key-challenge-for-foreign-companies/ [https://perma.cc/EX7C-H96E]. And, of course, technology companies are facing more aggressive antitrust enforcement in Europe as well as in the United States. *See* Adam Satariano & Matina Stevis-Gridneff, *Big Tech's Toughest Opponent Says She's Just Getting Started*, N.Y. Times (Nov. 19, 2019), https://www.nytimes.com/2019/11/19/technology/tech-regulator-europe.html [https://perma.cc/UVW3-AQYC].

[347] Daphne Keller (@daphnehk), Twitter (Apr. 5, 2018, 10:44 AM), https://twitter.com/daphnehk/status/981905519920103424 (sardonically labeling platforms' eagerness to demonstrate their policing capabilities as #VorauseilenderGehorsam, a German phrase meaning something along the lines of "preemptive obedience").

Conclusion

This Article has advanced three primary claims. First, the shift toward automated, *ex ante* content moderation was prompted and made possible by a legal architecture that insulated online intermediaries from liability and was intended to secure their independence. This architecture left space for intermediaries to develop new technologies and techniques that themselves became the law of moderation. Second, today, those same technologies are sites of contestation, cooptation, and increasing government control for user speech and privacy—reflecting the convergence of platform and government interests in surveillance and control. Finally, for these reasons, the modern regulations of online content that I have outlined are best understood not as challenges to platform power, but rather as reflections of platforms' own embeddedness in law enforcement, and vice versa.