THE NEW GOVERNORS: THE PEOPLE, RULES, AND PROCESSES GOVERNING ONLINE SPEECH

Kate Klonick

THE NEW GOVERNORS: THE PEOPLE, RULES, AND PROCESSES GOVERNING ONLINE SPEECH

Kate Klonick[*]

Private online platforms have an increasingly essential role in free speech and participation in democratic culture. But while it might appear that any internet user can publish freely and instantly online, many platforms actively curate the content posted by their users. How and why these platforms operate to moderate speech is largely opaque.

This Article provides the first analysis of what these platforms are actually doing to moderate online speech under a regulatory and First Amendment framework. Drawing from original interviews, archived materials, and internal documents, this Article describes how three major online platforms — Facebook, Twitter, and YouTube — moderate content and situates their moderation systems into a broader discussion of online governance and the evolution of free expression values in the private sphere. It reveals that private content-moderation systems curate user content with an eye to American free speech norms, corporate responsibility, and the economic necessity of creating an environment that reflects the expectations of their users. In order to accomplish this, platforms have developed a detailed system rooted in the American legal system with regularly revised rules, trained human decisionmaking, and reliance on a system of external influence.

This Article argues that to best understand online speech, we must abandon traditional doctrinal and regulatory analogies and understand these private content platforms as systems of governance. These platforms are now responsible for shaping and allowing participation in our new digital and democratic culture, yet they have little direct accountability to their users. Future intervention, if any, must take into account how and why these platforms regulate online speech in order to strike a balance between preserving the democratizing forces of the internet and protecting the generative power of our New Governors.

INTRODUCTION

In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we're really setting policies.

— Mark Zuckerberg[1]

---

[*] Ph.D. in Law Candidate, Yale University, and Resident Fellow at the Information Society at Yale Law School. Research for this project was made possible with the generous support of the Oscar M. Ruebhausen Fund. The author is grateful to Jack Balkin, Molly Brady, Kiel Brennan-Marquez, Peter Byrne, Adrian Chen, Bryan Choi, Danielle Keats Citron, Rebecca Crootof, Evelyn Frazee, Tarleton Gillespie, Eric Goldman, James Grimmelmann, Brad Greenberg, Alexandra Gutierrez, Woody Hartzog, David Hoffman, Gus Hurwitz, Thomas Kadri, Margot Kaminski, Alyssa King, Jonathan Manes, Toni Massaro, Christina Mulligan, Frank Pasquale, Robert Post, Sabeel Rahman, Jeff Rosen, Andrew Selbst, Jon Shea, Rebecca Tushnet, and Tom Tyler for helpful thoughts and comments on earlier versions of this Article. A special thank you to Rory Van Loo, whose own paper workshop inadvertently inspired me to pursue this topic. Elizabeth Goldberg and Deborah Won provided invaluable and brilliant work as research assistants.

[1] DAVID KIRKPATRICK, THE FACEBOOK EFFECT: THE INSIDE STORY OF THE COMPANY THAT IS CONNECTING THE WORLD 254 (2010).

In the summer of 2016, two historic events occurred almost simultaneously: a bystander captured a video of the police shooting of Alton Sterling on his cell phone, and another recorded the aftermath of the police shooting of Philando Castile and streamed the footage via Facebook Live.[2] Following the deaths of Sterling and Castile, Facebook founder and CEO Mark Zuckerberg stated that the ability to instantly post a video like the one of Castile dying "reminds us why coming together to build a more open and connected world is so important."[3] President Barack Obama issued a statement saying the shootings were "symptomatic of the broader challenges within our criminal justice system,"[4] and the Department of Justice opened an investigation into Sterling's shooting and announced that it would monitor the Castile investigation.[5] Multiple protests took place across the country.[6] The impact of these videos is an incredible example of how online platforms are now essential to participation in democratic culture.[7] But it almost never happened.

Initially lost in the voluminous media coverage of these events was a critical fact: as the video of Castile was streaming, it suddenly disappeared from Facebook.[8] A few hours later, the footage reappeared, this time with a label affixed warning of graphic content.[9] In official state-

---

[2] Richard Fausset et al., *Alton Sterling Shooting in Baton Rouge Prompts Justice Dept. Investigation*, N.Y. TIMES (July 6, 2016), http://nyti.ms/2szuH6H [https://perma.cc/Q8T6-HHXE]; Manny Fernandez et al., *11 Officers Shot, 4 Fatally, at Rally Against Violence*, N.Y. TIMES, July 8, 2016, at A1.

[3] Mark Zuckerberg, *Post*, FACEBOOK (July 7, 2016), https://www.facebook.com/zuck/posts/10102948714100101 [https://perma.cc/PP9M-FRTU].

[4] Press Release, White House, President Obama on the Fatal Shootings of Alton Sterling and Philando Castile (July 7, 2016), https://obamawhitehouse.archives.gov/blog/2016/07/07/president-obama-fatal-shootings-alton-sterling-and-philando-castile [https://perma.cc/VUL4-QT44].

[5] Press Release, U.S. Dep't of Justice, Attorney General Loretta E. Lynch Delivers Statement on Dallas Shooting (July 8, 2016), https://www.justice.gov/opa/speech/attorney-general-loretta-e-lynch-delivers-statement-dallas-shooting [https://perma.cc/UG89-XJXW].

[6] Fernandez et al., *supra* note 2; Liz Sawyer, *Protest Results in Brief Closure of State Fair's Main Gate*, STAR TRIB. (Minneapolis) (Sept. 3, 2016, 9:38 PM), http://www.startribune.com/protesters-gather-at-site-where-castile-was-shot/392247781/ [https://perma.cc/8Y4W-VF2Z]; Mitch Smith et al., *Peaceful Protests Follow Minnesota Governor's Call for Calm*, N.Y. TIMES (July 8, 2016), http://nyti.ms/2CqolWM [https://perma.cc/HRQ6-CTUC].

[7] *See* Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 2 (2004).

[8] *How Did Facebook Handle the Live Video of the Police Shooting of Philando Castile?*, WASH. POST (July 7, 2016, 11:45 AM), http://wapo.st/2ByazET [https://perma.cc/T6ZK-RZRC]; Mike Isaac & Sydney Ember, *Live Footage of Shootings Forces Facebook to Confront New Role*, N.Y. TIMES (July 8, 2016), http://nyti.ms/2CpLI37 [https://perma.cc/ZHR5-FAJE].

[9] Isaac & Ember, *supra* note 8.

ments, Facebook blamed the takedown on a "technical glitch" but provided no further details.[10] This is not entirely surprising. Though it might appear that any internet user can publish freely and instantly online, many content-publication platforms actively moderate[11] the content posted by their users.[12] Yet despite the essential nature of these platforms to modern free speech and democratic culture,[13] very little is known about how or why these companies curate user content.[14]

In response to calls for transparency, this Article examines precisely *what* these private platforms are actually doing to moderate user-generated content and *why* they are doing so. It argues that these platforms are best thought of as self-regulating[15] private entities, governing

---

[10] William Turton, *Facebook Stands by Technical Glitch Claim, Says Cop Didn't Delete Philando Castile Video*, GIZMODO (July 8, 2016, 1:36 PM), http://gizmodo.com/facebook-stands-by-technical-glitch-claim-says-cop-did-1783349993 [https://perma.cc/3ZWP-7SM9].

[11] I use the terms "moderate," "curate," and sometimes "regulate" to describe the behavior of these private platforms in both keeping up and taking down user-generated content. I use these terms rather than using the term "censor," which evokes the ideas of only removal of material and various practices of culturally expressive discipline or control. *See generally* Robert C. Post, *Project Report: Censorship and Silencing*, 51 BULL. AM. ACAD. ARTS & SCI. 32, 32 (1998). Where I do use "regulate," I do so in a more colloquial sense and not the way in which Professor Jack Balkin uses the term "speech regulation," which concerns government regulation of speech or government cooperation, coercion, or partnership with private entities to reflect government ends. *See* Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2299 (2014) (also explaining that the phrase "collateral censorship" is a term of art exempted from this taxonomy).

[12] *See* Catherine Buni & Soraya Chemaly, *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, THE VERGE (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech [https://perma.cc/PDM3-P6YH]; Adrian Chen, *The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed*, WIRED (Oct. 23, 2014, 6:30 AM), https://www.wired.com/2014/10/content-moderation/ [https://perma.cc/L5ME-T4H6]; Jeffrey Rosen, *Google's Gatekeepers*, N.Y. TIMES MAG. (Nov. 28, 2008), http://nyti.ms/2oc9lqw [https://perma.cc/YBM8-TNXC].

[13] Packingham v. North Carolina, 137 S. Ct. 1730, 1737 (2017) (holding that a state statute barring registered sex offenders from using online social media platforms was unconstitutional under the First Amendment). In his majority opinion, Justice Kennedy wrote that "[w]hile in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views, today the answer is clear. It is cyberspace — the 'vast democratic forums of the Internet' in general, and social media in particular." *Id.* at 1735 (citation omitted) (quoting Reno v. ACLU, 521 U.S. 844, 868 (1997)).

[14] *See, e.g.*, Marvin Ammori, *The "New"* New York Times*: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2273–76 (2014); Marjorie Heins, *The Brave New World of Social Media Censorship*, 127 HARV. L. REV. F. 325, 326 (2014) (describing Facebook's internal appeals process as "mysterious at best" and noting, about their internal policies, that "[t]he details of these rules . . . we do not know" and that the censorship "process in the private world of social media is secret").

[15] *See generally* Jody Freeman, *The Private Role in Public Governance*, 75 N.Y.U. L. REV. 543 (2000); Douglas C. Michael, *Federal Agency Use of Audited Self-Regulation as a Regulatory Technique*, 47 ADMIN. L. REV. 171 (1995).

speech within the coverage of the First Amendment[16] by reflecting the democratic culture and norms of their users.[17]

Part I surveys the regulatory and constitutional protections that have resulted in these private infrastructures. The ability of private platforms to moderate content comes from § 230 of the Communications Decency Act[18] (CDA), which gives online intermediaries broad immunity from liability for user-generated content posted on their sites.[19] The purpose of this grant of immunity was both to encourage platforms to be "Good Samaritans" and take an active role in removing offensive content, and also to avoid free speech problems of collateral censorship.[20] Beyond § 230, courts have struggled with how to conceptualize online platforms within First Amendment doctrine: as state actors, as broadcasters, or as editors. Additionally, scholars have moved between optimistic and pessimistic views of platforms and have long debated how — or whether — to constrain them.

To this legal framework and scholarly debate, this Article applies new evidence. Part II looks at why platforms moderate so intricately given the broad immunity of § 230. Through interviews with former platform architects and archived materials, this Article argues that platforms moderate content because of a foundation in American free speech norms, corporate responsibility, and the economic necessity of creating an environment that reflects the expectations of their users. Thus, platforms are motivated to moderate by both of § 230's purposes: fostering Good Samaritan platforms and promoting free speech.

Part III looks at how platforms are moderating user-generated content and whether that understanding can fit into an existing First Amendment framework. Through internal documents, archived materials, interviews with platform executives, and conversations with content moderators, this Article shows that platforms have developed a system that has marked similarities to legal or governance systems. This includes the creation of a detailed list of rules, trained human decisionmaking to apply those rules, and reliance on a system of external influence to update and amend those rules. With these facts, this Article

---

[16] *See generally* Balkin, *supra* note 11; Frederick Schauer, *The Boundaries of the First Amendment: A Preliminary Exploration of Constitutional Salience*, 117 HARV. L. REV. 1765 (2004).

[17] *See generally* ROBERT C. ELLICKSON, ORDER WITHOUT LAW (1991); ELINOR OSTROM, CRAFTING INSTITUTIONS FOR SELF-GOVERNING IRRIGATION SYSTEMS (1992); Balkin, *supra* note 7; J.M. Balkin, *Populism and Progressivism as Constitutional Categories*, 104 YALE L.J. 1935, 1948–49 (1995) (reviewing CASS R. SUNSTEIN, DEMOCRACY AND THE PROBLEM OF FREE SPEECH (1993), and defining democratic culture as popular participation in culture); Robert C. Ellickson, *Of Coase and Cattle: Dispute Resolution Among Neighbors in Shasta County*, 38 STAN. L. REV. 623 (1986).

[18] 47 U.S.C. § 230 (2012).

[19] *Id.*

[20] *See* Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (noting that the purposes of intermediary immunity in § 230 were not only to incentivize platforms to remove indecent content but also to protect the free speech of platform users).

argues that analogy under purely First Amendment doctrine should be largely abandoned.

Instead, platforms should be thought of as operating as the New Governors of online speech. These New Governors are part of a new triadic model of speech that sits between the state and speakers-publishers. They are private, self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free speech expectations of their users. Part IV explains how this conceptualization of online platforms as governance fits into scholarly concerns over the future of digital speech and democratic culture. It argues that the biggest threat this private system of governance poses to democratic culture is the loss of a fair opportunity to participate, which is compounded by the system's lack of direct accountability to its users. The first solution to this problem should not come from changes to § 230 or new interpretations of the First Amendment, but rather from simple changes to the architecture and governance systems put in place by these platforms. If this fails and regulation is needed, it should be designed to strike a balance between preserving the democratizing forces of the internet and protecting the generative power of our New Governors, with a full and accurate understanding of how and why these platforms operate, as presented here. It is only through accurately understanding the infrastructures and motivations of our New Governors that we can ensure that the free speech rights essential to our democratic culture remain protected.

I. SECTION 230, THE FIRST AMENDMENT, AND THE BEGINNINGS OF INTERMEDIARY SELF-REGULATION

Before the internet, the most significant constraint on the impact and power of speech was the publisher.[21] The internet ended the speaker's reliance on the publisher by allowing the speaker to reach his or her audience directly.[22] Over the last fifteen years, three American companies — YouTube, Facebook, and Twitter — have established themselves as dominant platforms in global content sharing.[23] These platforms are both the architecture for publishing new speech and the architects of the

---

[21] LAWRENCE LESSIG, CODE 2.0, at 19 (2006).

[22] *Id.*; Balkin, *supra* note 11, at 2306–10.

[23] *Facebook Grows as Dominant Content Sharing Destination*, MARKETING CHARTS (Aug. 24, 2016), https://www.marketingcharts.com/digital-70111 [https://perma.cc/VA4T-LM5Z] (describing Facebook and Twitter as the top content sharing destinations); *Facebook vs. YouTube: The Dominant Video Platform of 2017*, STARK CREW (Jan. 11, 2017), http://starkcrew.com/facebook-vs-youtube-the-dominant-video-platform-of-2017/ [https://perma.cc/5TTA-VJ64] (naming Facebook and YouTube as the dominant platforms for sharing video content online and summarizing their statistics).

institutional design that governs it.[24] This private architecture is the "central battleground over free speech in the digital era."[25]

A. History and Development of § 230

In order to understand the private governance systems used by platforms to regulate user content, it is necessary to start with the legal foundations and history that allowed for such a system to develop. The broad freedom of internet intermediaries[26] to shape online expression is based in § 230 of the CDA, which immunizes providers of "interactive computer services" from liability arising from user-generated content.[27] Sometimes called "the law that matters most for speech on the Web," the existence of § 230 and its interpretation by courts have been essential to the development of the internet as we know it today.[28]

Central to understanding the importance of § 230 are two cases decided before its existence, which suggested that intermediaries would be liable for defamation posted on their sites if they actively exercised any editorial discretion over offensive speech.[29] The first, *Cubby, Inc. v. CompuServe, Inc.*,[30] involved the publication of libel on CompuServe forums.[31] The court found CompuServe could not be held liable for the defamatory content in part because the intermediary did not review any of the content posted to the forum.[32] The *Cubby* court reasoned that CompuServe's practice of not actively reviewing content on its site made it more like a distributor of content, and not a publisher.[33] In determining communication tort liability, this distinction is important because while publishers and speakers of content can be held liable, distributors are generally not liable unless they knew or should have known of the

---

[24] LESSIG, *supra* note 21, at 2–10 (describing the internet as architecture).

[25] Balkin, *supra* note 11, at 2296.

[26] Internet intermediaries are broadly defined as actors in every part of the internet "stack." *See* JAMES GRIMMELMANN, INTERNET LAW 31 (2016). These include internet service providers, hosting providers, servers, websites, social networks, search engines, and so forth. *See id.* at 31–32. Within this array, I use "platforms" to refer specifically to internet websites or apps that publish user content — these include Facebook, YouTube, and Twitter.

[27] 47 U.S.C. § 230(c)(2) (2012); *see also* Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (blocking claims against AOL under § 230 because AOL was only the publisher, and not the creator, of the tortious content).

[28] Emily Bazelon, *How to Unmask the Internet's Vilest Characters*, N.Y. TIMES MAG. (Apr. 22, 2011), http://nyti.ms/2C30ZL9 [https://perma.cc/55A3-6FAN].

[29] *See* Davis S. Ardia, *Free Speech Savior or Shield for Scoundrels: An Empirical Study of Intermediary Immunity Under Section 230 of the Communications Decency Act*, 43 LOY. L.A. L. REV. 373, 406–09 (2010).

[30] 776 F. Supp. 135 (S.D.N.Y. 1991).

[31] *Id.* at 138; Ardia, *supra* note 29, at 406–07. CompuServe did not dispute that the statements were defamatory. *Cubby*, 776 F. Supp. at 138.

[32] *Cubby*, 776 F. Supp. at 140.

[33] *Id.* at 139–41.

defamation.[34] Though distributor-publisher distinctions were an established analogy in tort liability, the difficulty of using this model for online intermediaries quickly became apparent. Four years after *Cubby*, in *Stratton Oakmont, Inc. v. Prodigy Services Co.*,[35] a court found that the intermediary Prodigy was liable as a publisher for all posts made on its site because it actively deleted some forum postings.[36] To many, Prodigy's actions seemed indistinguishable from those that had rendered CompuServe a mere distributor in *Cubby*, but the court found Prodigy's use of automatic software and guidelines for posting were a "conscious choice, to gain the benefits of editorial control."[37] Read together, the cases seemed to expose intermediaries to a wide and unpredictable range of tort liability if they exercised any editorial discretion over content posted on their sites. Accordingly, the cases created a strong disincentive for online intermediaries to expand business or moderate offensive content and threatened the developing landscape of the internet.

Thankfully, the developing landscape of the internet was an active agenda item for Congress when the *Stratton Oakmont* decision came down. Earlier that year, Senator James Exon had introduced the CDA, which aimed to regulate obscenity online by making it illegal to knowingly send or show minors indecent online content.[38] Reacting to the concerns created by *Stratton Oakmont*, Representatives Chris Cox and Ron Wyden introduced an amendment to the CDA that would become § 230.[39] The Act, with the Cox-Wyden amendment, passed and was signed into law in February 1996.[40] In its final form, § 230(c) stated that "[n]o 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"[41] in order to incentivize and protect intermediaries' Good Samaritan blocking of offensive material.[42] Though, just a little over a year later, the Supreme Court in *Reno v. ACLU*[43] struck down the bulk of the anti-indecency sections of the CDA, § 230 survived.[44]

---

[34] RESTATEMENT (SECOND) OF TORTS § 581(1) (AM. LAW INST. 1977).

[35] 1995 WL 323710 (N.Y. Sup. Ct. May 24, 1995).

[36] *Id.* at *4.

[37] *Id.* at *5.

[38] *See* Robert Cannon, *The Legislative History of Senator Exon's Communications Decency Act: Regulating Barbarians on the Information Superhighway*, 49 FED. COMM. L.J. 52–53 (1996).

[39] 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statements of Reps. Cox, Wyden, and Barton).

[40] *See* Pub. L. No. 104-104, tit. V, 110 Stat. 56, 133–43 (1996) (codified in scattered sections of 18 and 47 U.S.C.); *see also* H.R. REP. NO. 104-458, at 81–91 (1996); S. REP. NO. 104-230, at 187–93 (1996); S. REP. NO. 104-23, at 9 (1995). For a full and thorough account of the legislative history of § 230, see generally Cannon, *supra* note 38.

[41] 47 U.S.C. § 230(c)(1) (2012).

[42] 141 CONG. REC. H8469–70 (statement of Rep. Cox).

[43] 521 U.S. 844 (1997).

[44] *Id.* at 885.

It soon became clear that § 230 would do more than just survive. A few months after *Reno*, the Fourth Circuit established a foundational and expansive interpretation of § 230 in *Zeran v. America Online, Inc.*[45] Plaintiff Zeran sought to hold AOL liable for defamatory statements posted on an AOL message board by a third party.[46] Zeran argued that AOL had a duty to remove the posting, post notice of the removed post's falsity, and screen future defamatory material.[47] The court disagreed. Instead, it found AOL immune under § 230 and held that the section precluded not only strict liability for publishers but also intermediary liability for distributors such as website operators.[48] This holding also extinguished notice liability for online intermediaries.[49]

While the holdings in *Zeran* were broad and sometimes controversial,[50] it was the court's analysis as to the purposes and scope of § 230 that truly shaped the doctrine. In granting AOL the affirmative defense of immunity under § 230, the court recognized the Good Samaritan provision's purpose of encouraging "service providers to self-regulate the dissemination of offensive material over their services."[51] But the court did not consider § 230 merely a congressional response to *Stratton Oakmont*. Instead, the court looked to the plain text of § 230(c) granting statutory immunity to online intermediaries and drew new purpose beyond the Good Samaritan provision and found that intent "not difficult to discern":

> Congress recognized the threat that tort-based lawsuits pose to *freedom of speech* in the new and burgeoning Internet medium. The imposition of tort liability on service providers for the communications of others represented, for Congress, simply another form of *intrusive government regulation of speech*.[52]

Thus, while the court reasoned that § 230 lifted the "specter of tort liability" that might "deter service providers from blocking and screening offensive material," it found it was also Congress's design to immunize intermediaries from any requirement to do so.[53] Drawing on these free speech concerns, the court reasoned that the same "specter of tort liability" that discouraged intermediaries from policing harmful content also threatened "an area of such prolific speech" with "an obvious

---

[45] 129 F.3d 327 (4th Cir. 1997).

[46] *Id.* at 328.

[47] *Id.* at 330.

[48] *Id.* at 332.

[49] *Id.* at 333.

[50] *See Developments in the Law — The Law of Cyberspace*, 112 HARV. L. REV. 1574, 1613 (1999) (referring to *Zeran*'s holding as a "broad interpretation of § 230").

[51] *Zeran*, 129 F.3d at 331.

[52] *Id.* at 330 (emphases added).

[53] *Id.* at 331.

chilling effect."[54] "Faced with potential liability for each message republished by their services, interactive computer service providers might choose to severely restrict the number and type of messages posted."[55] In response to the question raised by Zeran of subjecting publishers like AOL to notice-based liability, the court again cited its free speech concerns but also recognized the practical realities of distributors: "Each notification would require a careful yet rapid investigation of the circumstances surrounding the posted information, a legal judgment concerning the information's defamatory character, and an on-the-spot editorial decision whether to risk liability by allowing the continued publication of that information."[56]

The sheer volume of content to be policed by intermediaries, and their almost certain liability should they be notified and still publish, would lead to either haphazard takedowns at best, or widespread removal at worst. "Thus, like strict liability, liability upon notice has a chilling effect on the freedom of Internet speech."[57]

Zeran is a seminal decision in internet law not only because it gave broad immunity to online intermediaries[58] but also because of its analysis of the purposes of § 230. The court recognized two distinct congressional purposes for granting immunity under § 230: (1) as a Good Samaritan provision written to overturn *Stratton Oakmont* and "to encourage interactive computer services and users of such services to self-police the Internet for obscenity and other offensive material,"[59] and (2)

---

[54] *Id.*

[55] *Id.* The quote continues: "Congress considered the weight of the speech interests implicated and chose to immunize service providers to avoid any such restrictive effect." *Id.*

[56] *Id*. at 333.

[57] *Id*. Though this free speech purpose might not have been in the plain text of § 230, the *Zeran* court did not invent it. *See* Cannon, *supra* note 38, at 88–91 (discussing the legislative history indicating that Congress debated the "contest between censorship and democratic discourse," *id.* at 88).

[58] A number of scholars have criticized the reasoning in *Zeran* and its progeny for this reason. *See, e.g.*, Susan Freiwald, *Comparative Institutional Analysis in Cyberspace: The Case of Intermediary Liability for Defamation*, 14 HARV. J.L. & TECH. 569, 594–96 (2001); Sewali K. Patel, *Immunizing Internet Service Providers from Third-Party Internet Defamation Claims: How Far Should Courts Go?*, 55 VAND. L. REV. 647, 679–89 (2002); David R. Sheridan, Zeran v. AOL *and the Effect of Section 230 of the Communications Decency Act upon Liability for Defamation on the Internet*, 61 ALB. L. REV. 147, 169–70 (1997); Michael H. Spencer, *Defamatory E-Mail and Employer Liability: Why Razing* Zeran v. America Online *Is a Good Thing*, 6 RICH. J.L. & TECH. 25 (2000); Michelle J. Kane, Note, Blumenthal v. Drudge, 14 BERKELEY TECH. L.J. 483, 498–500 (1999); Brian C. McManus, Note, *Rethinking Defamation Liability for Internet Service Providers*, 35 SUFFOLK U. L. REV. 647, 667–68 (2001); Annemarie Pantazis, Note, Zeran v. America Online, Inc.*: Insulating Internet Service Providers from Defamation Liability*, 34 WAKE FOREST L. REV. 531, 547–50 (1999); David Wiener, Note, *Negligent Publication of Statements Posted on Electronic Bulletin Boards: Is There* Any *Liability Left After* Zeran*?*, 39 SANTA CLARA L. REV. 905 (1999).

[59] Batzel v. Smith, 333 F.3d 1018, 1028 (9th Cir. 2003) (first citing 47 U.S.C. § 230(b)(4) (2012); then citing 141 CONG. REC. H8469–70 (daily ed. Aug. 4, 1995) (statements of Reps. Cox, Wyden,

as a free speech protection for users meant "to encourage the unfettered and unregulated development of free speech on the Internet, and to promote the development of e-commerce."[60]

Though the exact term is not stated in the text of *Zeran*, the court's concern over service providers' "natural incentive simply to remove messages upon notification, whether the contents were defamatory or not," reflects apprehension of collateral censorship.[61] Collateral censorship occurs when one private party, like Facebook, has the power to control speech by another private party, like a Facebook user.[62] Thus, if the government threatens to hold Facebook liable based on what its user says, and Facebook accordingly censors its user's speech to avoid liability, you have collateral censorship.[63] The court in *Zeran* recognized this concern for the free speech rights of users and counted it among the reasons for creating immunity for platforms under § 230.

But while the dual purposes of § 230 call for the same solution — intermediary immunity — they create a paradox in the applications of § 230. If § 230 can be characterized as both government-created immunity to (1) *encourage* platforms to remove certain kinds of content, and (2) *avoid* the haphazard removal of certain content and the perils of collateral censorship to users, which interests do we want to prioritize? That of the platforms to moderate their content or that of users' free speech?

In the last few years, courts have grappled with precisely this dilemma and occasionally broken with the expansive interpretation of the Good Samaritan provision to find a lack of § 230 immunity.[64] For instance, in two recent district court cases in northern California, the court

---

and Barton); then citing *Zeran*, 129 F.3d at 331; and then citing Blumenthal v. Drudge, 992 F. Supp. 44, 52 (D.D.C. 1998)).

60 *Id.* at 1027–28 (first citing § 230(b) (policy objectives include "(1) to promote the continued development of the Internet and other interactive computer services and other interactive media; (2) to preserve the vibrant and competitive free market that presently exists for the Internet and other interactive computer services, unfettered by Federal or State regulation"); then citing *Zeran*, 129 F.3d at 330).

61 *Zeran*, 129 F.3d at 333. The court also specifically cited worry about potential abuse between users. "Whenever one was displeased with the speech of another party conducted [online], the offended party could simply 'notify' the relevant service provider, claiming the information to be legally defamatory." *Id.*; *see also* Christina Mulligan, *Technological Intermediaries and Freedom of the Press*, 66 SMU L. REV. 157, 171 (2013); Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293, 317–18 (2011).

62 The term "collateral censorship" was coined by Professor Michael Meyerson. Michael I. Meyerson, *Authors, Editors, and Uncommon Carriers: Identifying the "Speaker" Within the New Media*, 71 NOTRE DAME L. REV. 79, 118 (1995).

63 *Cf.* J.M. Balkin, Essay, *Free Speech and Hostile Environments*, 99 COLUM. L. REV. 2295, 2298 (1999).

64 For a comprehensive cataloging of § 230 cases with context and commentary, see Eric Goldman, *Ten Worst Section 230 Rulings of 2016 (Plus the Five Best)*, TECH. & MARKETING L.

rejected motions to dismiss for failure to state a claim under § 230 on the basis of plaintiffs' allegations that Google acted in bad faith.[65] At the same time, other courts have made powerful decisions in favor of broad § 230 immunity and publishers' rights to moderate content. Notably, in *Doe v. Backpage.com*,[66] the First Circuit expressly held that § 230 protects the choices of websites as speakers and publishers, stating: "Congress did not sound an uncertain trumpet when it enacted the CDA, and it chose to grant broad protections to internet publishers. Showing that a website operates through a meretricious business model is not enough to strip away those protections."[67] The continued confusion about § 230's interpretation — as seen in current courts' split on the importance of a business's motivations for content moderation — demonstrates that the stakes around such questions have only grown since the foundational decision in *Zeran*.

B. First Amendment Implications

The debate over how to balance the right of intermediaries to curate a platform while simultaneously protecting user speech under the First Amendment is ongoing for courts and scholars. Depending on the type of intermediary involved, courts have analogized platforms to established doctrinal areas in First Amendment law — company towns, broadcasters, editors — and the rights and obligations of a platform shift depending on which analogy is applied.

The first of these analogies reasons that platforms are acting like the state, so the First Amendment directly constrains them. While courts have established that only state action creates affirmative obligations under the First Amendment, determining exactly when a private party's behavior constitutes state action is a more difficult question.[68] The Supreme Court foundationally addressed this distinction between private and state actors for First Amendment purposes in *Marsh v. Alabama*.[69] In *Marsh*, a Jehovah's Witness was arrested for criminal trespass for distributing literature on the sidewalk of a company town

---

BLOG (Jan. 4, 2017), http://blog.ericgoldman.org/archives/2017/01/ten-worst-section-230-rulings-of-2016-plus-the-five-best.htm [https://perma.cc/KL48-B6GJ].

65 Darnaa, LLC v. Google, Inc., No. 15-cv-03221, 2016 WL 6540452 (N.D. Cal. Nov. 2, 2016); Spy Phone Labs LLC v. Google Inc., No. 15-cv-03756, 2016 WL 6025469 (N.D. Cal. Oct. 14, 2016); *see also* Eric Goldman, *Google Loses Two Section 230(c)(2) Rulings* — Spy Phone v. Google *and* Darnaa v. Google, TECH. & MARKETING L. BLOG (Nov. 8, 2016), http://blog.ericgoldman.org/archives/2016/11/google-loses-two-section-230c2-rulings-spy-phone-v-google-and-darnaa-v-google.htm [https://perma.cc/TR72-9XZU].

66 817 F.3d 12 (1st Cir. 2016).

67 *Id.* at 29.

68 *See* Hudgens v. NLRB, 424 U.S. 507, 513–21 (1976).

69 326 U.S. 501 (1946).

wholly owned by a corporation.[70] The Court found that "[e]xcept for [ownership by a private corporation, this town] has all the characteristics of any other American town."[71] Accordingly, the Court held the town was functionally equivalent to a state actor and obligated to guarantee First Amendment rights.[72]

In the years since *Marsh*, the Court has continued to explore the "public function" circumstances necessary for private property to be treated as public. Many of these cases have arisen in the context of shopping malls, where the Court has struggled to establish consistent reasoning on when a private individual's First Amendment rights trump the rights of the owner of a private forum.[73] The most expansive of these was *Amalgamated Food Employees Union Local 590 v. Logan Valley Plaza, Inc.*,[74] which held a shopping mall to be the equivalent of the company town in *Marsh* and therefore allowed picketers to protest there.[75] In overruling *Logan Valley* in *Hudgens v. NLRB*,[76] the Court revised its assessment of a shopping mall as a public square and stated that a business does not qualify as performing a public function merely because it is open to the public.[77] Instead, in order to qualify as performing a public function, a business must be actually doing a job normally done by the government, as was the case with the company town in *Marsh*.[78]

For a long time, the claim that online intermediaries are state actors or perform a public function and, thus, are subject to providing free speech guarantees, was a losing one. In establishing platforms as non-state actors, courts distinguished the facts in *Marsh* and its progeny, stating that intermediaries providing services like email, hosting, or search engines do not rise to the level of "performing any municipal power or essential public service and, therefore, do[] not stand in the

---

[70] *Id.* at 502–03.

[71] *Id.* at 502.

[72] *Id.* at 508–09.

[73] *See, e.g.*, Amalgamated Food Emps. Union Local 590 v. Logan Valley Plaza, Inc., 391 U.S. 308, 318 (1968) (equating a private shopping center to a business district and affirming the right to picket in it), *narrowed by* Lloyd Corp. v. Tanner, 407 U.S. 551, 563–64 (1972) (holding speech in a mall is not constitutionally protected unless there are no other means of communication), *overruled by Hudgens*, 424 U.S. at 518. The California Supreme Court granted more expansive free speech guarantees than those provided by the First Amendment in *Fashion Valley Mall, LLC v. NLRB*, 172 P.3d 742, 749 (Cal. 2007), and *Robins v. PruneYard Shopping Center*, 592 P.2d 341, 344, 347 (Cal. 1979). *See also Developments in the Law — State Action and the Public/Private Distinction*, 123 HARV. L. REV. 1248, 1303–07 (2010).

[74] 391 U.S. 308.

[75] *Id.* at 318.

[76] 424 U.S. 507.

[77] *Id.* at 519 (quoting *Lloyd Corp.*, 407 U.S. at 568–69).

[78] *Id.*

shoes of the State."[79] While these cases have not been explicitly overturned, the Court's recent ruling in *Packingham v. North Carolina*[80] might breathe new life into the application of state action doctrine to internet platforms.

In *Packingham*, the Court struck down a North Carolina statute barring registered sex offenders from platforms like Facebook and Twitter.[81] In his opinion for the court, Justice Kennedy reasoned that foreclosing "access to social media altogether is to prevent the user from engaging in the legitimate exercise of First Amendment rights."[82] Describing such services as a "modern public square," Justice Kennedy also acknowledged their essential nature to speech, calling them "perhaps the most powerful mechanisms available to a private citizen to make his or her voice heard."[83] Though the decision is limited in that it applies only to total exclusion, the sweeping language makes *access* to private online platforms a First Amendment right, leaving open the questions of how robust that access must be or where in the internet pipeline a choke point must lie in order to abridge a First Amendment right. Future litigation might use *Packingham*'s acknowledgment of a First Amendment right to social media access as a new basis to argue that these platforms perform quasi-municipal functions.

Separate from the issue of state action, *Packingham*'s acknowledgment of platforms as private forums that significantly affect the expressive conduct of other private parties implicates other areas of regulation that are consistent with the First Amendment. This can be seen in the doctrine around other types of speech conduits, like radio and television broadcasters. In such cases, the Court has upheld regulation of radio broadcasting, despite the broadcast station's claims that the regulation unconstitutionally infringed on its editorial judgment and speech.[84] A public right to "suitable access" to ideas and a scarce radio spectrum justified the agency rule that required broadcasters to present public

---

[79] Cyber Promotions, Inc. v. Am. Online, Inc., 948 F. Supp. 436, 442 (E.D. Pa. 1996) (distinguishing AOL's email service from the kind of "municipal powers or public services" provided by a private company town that made it liable as a state actor in *Marsh*); *see also* Green v. Am. Online, 318 F.3d 465, 472 (3d Cir. 2003) (holding that, as a private company and not a state actor, AOL is not subject to constitutional free speech requirements); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 631 (D. Del. 2007) (finding that for the purposes of constitutional free speech guarantees, Google, Yahoo, and Microsoft are private companies, even though they work with state actors like public universities).

[80] 137 S. Ct. 1730 (2017).

[81] *Id.* at 1733, 1738.

[82] *Id.* at 1737.

[83] *Id.*

[84] *See, e.g.*, Red Lion Broad. Co. v. FCC, 395 U.S. 367 (1969).

issues and give each side of those issues fair coverage.[85] In the years following, the Court has limited this holding,[86] while also extending it to the realm of broadcast television in *Turner Broadcasting System, Inc. v. FCC*.[87]

The question of whether internet intermediaries would fall in the same category as radio or broadcast television was addressed by the Court in *Reno*. The Court found that the elements that justify television and radio regulation — those mediums' "invasive" nature, history of extensive regulation, and the scarcity of frequencies — "are not present in cyberspace" and explicitly exempted the internet from the doctrine established in *Red Lion Broadcasting Co. v. FCC*[88] and *Turner*.[89] While it is unclear how the Court would draw the line between the internet functions of concern in *Reno* and the growth of social media platforms, *Packingham*'s emphasis on the right to platform access might revive the concerns over scarcity raised by these cases.

The final First Amendment analogy relevant to online speech reasons that platforms themselves exercise an important expressive role in the world, so the First Amendment actively protects them from state interference. This draws on the doctrine giving special First Amendment protections to newspapers under *Miami Herald Publishing Co. v. Tornillo*.[90] There, in a unanimous decision, the Court found a Florida statute that gave political candidates a "right to reply" in local newspapers unconstitutional under the Free Press Clause of the First Amendment.[91] Though the "right to reply" legislation was akin to FCC fairness regulations upheld in *Red Lion*, the *Tornillo* Court found the statute unconstitutional.[92] The Court reasoned that the statute was an

---

[85] *Id.* at 400–01 ("In view of the scarcity of broadcast frequencies, the Government's role in allocating those frequencies, and the legitimate claims of those unable without governmental assistance to gain access to those frequencies for expression of their views, we hold the regulations and ruling at issue here are both authorized by statute and constitutional.").

[86] *See, e.g.*, FCC v. League of Women Voters, 468 U.S. 364, 402 (1984) (holding publicly funded broadcasters have First Amendment protections to editorialize); FCC v. Pacifica Found., 438 U.S. 726, 741 n.17 (1978) (stating "it is well settled that the First Amendment has a special meaning in the broadcasting context" and citing *Red Lion*); Columbia Broad. Sys., Inc. v. Democratic Nat'l Comm., 412 U.S. 94, 120–21 (1973) (holding broadcasters are not under an obligation to sell advertising time to a political party).

[87] Turner Broad. Sys., Inc. v. FCC (*Turner II*), 520 U.S. 180, 185 (1997); Turner Broad. Sys., Inc. v. FCC (*Turner I*), 512 U.S. 622, 638–39 (1994). In these cases the Court dealt with FCC "must carry" regulations imposed on cable television companies. In *Turner I*, the Court determined that cable television companies were indeed First Amendment speakers, 512 U.S. at 656, but in *Turner II*, it held that the "must carry" provisions of the FCC did not violate those rights, 520 U.S. at 224–25.

[88] 395 U.S. 367.

[89] Reno v. ACLU, 521 U.S. 844, 868–70 (1997).

[90] 418 U.S. 241 (1974).

[91] *Id.* at 247, 258.

[92] *Id.* at 258.

"intrusion into the function of editors"[93] and that "press responsibility is not mandated by the Constitution and . . . cannot be legislated."[94] As internet intermediaries have become more and more vital to speech, First Amendment advocates have urged courts to apply the holding in *Tornillo* to platforms, granting them their own speech rights.[95] The Court's new definition in *Packingham* of online speech platforms as forums, however, might threaten the viability of arguments that these companies have their own First Amendment rights as speakers.

C. Internet Pessimists, Optimists, and Realists

As have the courts, scholars have struggled with the question of how to balance users' First Amendment right to speech against intermediaries' right to curate platforms. Many look to platforms as a new market for speech and ideas. In the early days of the internet, Professor Jack Balkin could have been considered an internet optimist. He saw the internet and its wealth of publishing tools, which enable widespread digital speech, as enhancing the "possibility of democratic culture."[96] More recently, he has recognized that private control of these tools poses threats to free speech and democracy.[97] Professor Yochai Benkler could also have been considered an optimist, though a more cautious one. He has posited looking at the internet as enabling new methods of information production, as well as a move from traditional industrial-dominated markets to more collaborative peer production.[98] Professor Lawrence Lessig acknowledges that while the internet creates exciting new means to regulate through code, he is concerned about corporations and platforms having great unchecked power to regulate the internet and all interactions that fall under § 230 immunity.[99] Professors James Boyle, Jack Goldsmith, and Tim Wu have had similar concerns about

---

[93] *Id.*

[94] *Id.* at 256.

[95] *See* Eric Goldman, *Revisiting Search Engine Bias*, 38 WM. MITCHELL L. REV. 96, 108–10 (2011); Eugene Volokh & Donald M. Falk, *First Amendment Protection for Search Engine Search Results*, VOLOKH CONSPIRACY (Apr. 20, 2012), http://www.volokh.com/wp-content/uploads/2012/05/SearchEngineFirstAmendment.pdf [https://perma.cc/U27F-MA6U]. *But see* James Grimmelmann, *Some Skepticism About Search Neutrality*, *in* THE NEXT DIGITAL DECADE: ESSAYS ON THE FUTURE OF THE INTERNET 435 (Berin Szoka & Adam Marcus eds., 2010); Frank Pasquale, *Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power*, 17 THEORETICAL INQUIRIES L. 487, 502–03 (2016) (refuting efforts to apply *Tornillo* to internet intermediaries).

[96] Balkin, *supra* note 7, at 45–46.

[97] *See* Balkin, *supra* note 11, at 2300–01.

[98] *See generally* YOCHAI BENKLER, THE WEALTH OF NETWORKS (2006); Yochai Benkler, *Through the Looking Glass: Alice and the Constitutional Foundations of the Public Domain*, 66 LAW & CONTEMP. PROBS. 173, 181–82 (2003).

[99] *See generally* LESSIG, *supra* note 21; Lawrence Lessig, Commentary, *The Law of the Horse: What Cyberlaw Might Teach*, 113 HARV. L. REV. 501 (1999).

the state coopting private online intermediaries for enforcement.[100] Professor David Post has argued that the market will resolve corporate monopolization of speech. He has suggested that such corporate competition between individual online platforms would result in a "market for rules," which would allow users to seek networks that have speech and conduct "rule sets" to their liking.[101]

Not quite optimists or pessimists, many internet scholars have focused their work on the realities of what the internet is, the harms it does and can create, and the best ways to resolve those harms. Professor Danielle Keats Citron was an early advocate for this approach. She has argued for recognition of cyber civil rights in order to circumvent § 230 immunity without removing the benefits of its protection.[102] Professor Mary Anne Franks has continued this tack, and argues that the nature of online space can amplify speech harms, especially in the context of sexual harassment.[103] Online hate speech, harassment, bullying, and revenge porn have slightly different solutions within these models. Both Citron and Professor Helen Norton have argued that hate speech is now mainstream and should be actively addressed by platforms that have the most power to curtail it.[104] Emily Bazelon argues that the rise of online bullying calls for a more narrow reading of § 230.[105] Citron and Franks respectively suggest either an amendment or a court-created narrowing of § 230 for sites that host revenge porn.[106]

This is where we stand today in understanding internet intermediaries: amidst a § 230 dilemma (is it about enabling platforms to edit their sites or about protecting users from collateral censorship?), a First Amendment enigma (what are online platforms for the purposes of speech — a company town, a broadcaster, or an editor?), and conflicting scholarly theories of how best to understand speech on the internet.

---

100 *See generally* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? (2006); James Boyle, *Foucault in Cyberspace: Surveillance, Sovereignty, and Hardwired Censors*, 66 U. CIN. L. REV. 177 (1997); *see also* Rory Van Loo, *Rise of the Digital Regulator*, 66 DUKE L.J. 1267, 1267 (2017) (discussing how the state is using online platforms to enforce consumer protection and generally regulate markets in place of legal rules).

101 David G. Post, *Anarchy, State, and the Internet: An Essay on Law-Making in Cyberspace*, 1995 J. ONLINE L. art. 3, para. 42. *But see* Frank Pasquale, *Privacy, Antitrust, and Power*, 20 GEO. MASON L. REV. 1009 (2013) (arguing that platforms like Facebook, Twitter, LinkedIn, and Instagram are complements, not substitutes, for one another).

102 *See* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE (2014); Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 115–25 (2009).

103 Mary Anne Franks, *Sexual Harassment 2.0*, 71 MD. L. REV. 655, 678, 681–83 (2012).

104 Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. REV. 1435, 1456–68 (2011).

105 *See generally* EMILY BAZELON, STICKS AND STONES: DEFEATING THE CULTURE OF BULLYING AND REDISCOVERING THE POWER OF CHARACTER AND EMPATHY (2013); Bazelon, *supra* note 28.

106 Danielle Keats Citron & Mary Anne Franks, *Criminalizing Revenge Porn*, 49 WAKE FOREST L. REV. 345, 359 n.86 (2014).

Missing from the debate around § 230 is the answer to a simple question: given that these platforms have § 230 immunity, *why* are they bothering to edit? Administrative law scholarship discusses the forces that motivate private actors to voluntarily self-regulate.[107] Some firms or industries have developed self-regulation alongside government regulation.[108] Others see self-regulation as an optimal form of business and company management.[109] And some decide to self-regulate as an attempt to preempt eventual government regulation.[110] Some of these reasons come to bear on platform motivation, but because of immunity under § 230, most are irrelevant. Instead, through historical interviews and archived materials, Part II argues that platforms have created a voluntary system of self-regulation because they are economically motivated to create a hospitable environment for their users in order to incentivize engagement.[111] This self-regulation involves both reflecting the norms of their users around speech as well as keeping up as much speech as possible. Online platforms also self-regulate for reasons of social and corporate responsibility, which in turn reflect free speech norms.[112] These motivations reflect both the Good Samaritan incentives and collateral censorship concerns underlying § 230.

A question is also missing from the debate about how to classify platforms in terms of First Amendment doctrine: what are major online intermediaries *actually doing* to regulate content on their sites? The next Part discusses *why* platforms are making the decisions to moderate along such a fine line, while the following Part demonstrates *how* platforms moderate content through a detailed set of rules, trained human decisionmaking, and reasoning by analogy, all influenced by a pluralistic system of internal and external actors.

---

107 *See* Freeman, *supra* note 15, at 644–49; Michael, *supra* note 15, at 203–40.

108 *See, e.g.*, JOSEPH V. REES, HOSTAGES OF EACH OTHER: THE TRANSFORMATION OF NUCLEAR SAFETY SINCE THREE MILE ISLAND 1–2 (1994) (documenting private use of self-regulation in an industrial area following disaster).

109 *See generally* DENNIS C. KINLAW, CONTINUOUS IMPROVEMENT AND MEASUREMENT FOR TOTAL QUALITY (1992) (describing self-regulation, specifically through the use of total quality management and self-auditing, as the best technique for business management and means of achieving customer satisfaction).

110 *See* RICHARD L. ABEL, AMERICAN LAWYERS 142–57 (1989) (discussing private actors' decisions to self-regulate in order to avoid potential government regulation).

111 *See* Citron & Norton, *supra* note 104, at 1454 (discussing how some intermediaries regulate hate speech because they see it as a threat to profits).

112 *Id.* at 1455 (discussing how some intermediaries regulate hate speech because they see it as a corporate or social responsibility).

II. WHY GOVERN WELL? THE ROLE OF FREE SPEECH NORMS, CORPORATE CULTURE, AND ECONOMIC INCENTIVES IN THE DEVELOPMENT OF CONTENT MODERATION

In the earliest days of the internet, the regulations concerning the substance and structure of cyberspace were "built by a noncommercial sector [of] researchers and hackers, focused upon building a network."[113] Advances in technology as well as the immunity created for internet intermediaries under § 230 led to a new generation of cyberspace. It included collaborative public platforms like Wikipedia,[114] but it was also populated largely by private commercial platforms.[115]

As this online space developed, scholars considered what normative values were being built into the infrastructure of the internet. Lessig ascribed a constitutional architecture to the internet "not to describe a hundred-day plan[, but] instead to identify the values that a space should guarantee. . . . [W]e are simply asking: What values should be protected there? What values should be built into the space to encourage what forms of life?"[116] Writing five years later in 2004,[117] Balkin argued that the values of cyberspace are inherently democratic — bolstered by the ideals of free speech, individual liberty, and participation.[118] Both Lessig and Balkin placed the fate of "free speech values"[119] and the "freedoms and controls of cyberspace"[120] in the hands of code and architecture online.[121] "[A] code of cyberspace, defining the freedoms and controls of cyberspace, will be built," wrote Lessig.[122] "About that there can be no debate. But by whom, and with what values? That is the only choice we have left to make."[123]

There was not much choice about it, but over the last fifteen years, three American companies — YouTube, Facebook, and Twitter — have

---

113 LESSIG, *supra* note 21, at 7.

114 *See generally* Yochai Benkler, *Yochai Benkler on Wikipedia's 10th Anniversary*, THE ATLANTIC (Jan. 15, 2011), https://www.theatlantic.com/technology/archive/2011/01/yochai-benkler-on-wikipedias-10th-anniversary/69642/ [https://perma.cc/2W32-4EFV].

115 LESSIG, *supra* note 21, at 7 (describing the second generation of the internet as being "built by commerce").

116 *Id.* at 6.

117 As calculated from the first distribution of Lessig's book, LAWRENCE LESSIG, CODE AND OTHER LAWS OF CYBERSPACE (1999).

118 *See* Balkin, *supra* note 7, at 45–49.

119 *Id.* at 54.

120 LESSIG, *supra* note 21, at 6.

121 Specifically, Balkin predicted that free speech values of "participation, access, interactivity, democratic control, and the ability to route around and glom on . . . won't necessarily be protected and enforced through judicial creation of constitutional rights. Rather, they will be protected and enforced through the design of technological systems — code — and through legislative and administrative schemes of regulation." Balkin, *supra* note 7, at 54.

122 LESSIG, *supra* note 21, at 6.

123 *Id.*

established themselves as dominant platforms in global content sharing and online speech.[124] These platforms are both the architecture for publishing new speech and the architects of the institutional design that governs it. Because of the wide immunity granted by § 230, these architects are free to choose which values they want to protect — or to protect no values at all. So why have they chosen to integrate values into their platform? And what values have been integrated?

It might first be useful to describe what governance means in the context of these platforms. "The term 'governance' is popular but imprecise," and modern use does not assume "governance as a synonym for government."[125] Rather, "new governance model[s]" identify several features that accurately describe the interplay between user and platform: a "dynamic" and "iterative" "law-making process";[126] "norm-generating" "[i]ndividuals";[127] and "convergence of processes and outcomes."[128] This is the way in which this Article uses the term "governance." However, the user-platform relationship departs from even this definition because of its private and centralized but also pluralistically networked nature. And it departs even further from other uses of the term "governance," including "corporate governance" (describing it as centralized management) and public service definitions of "good governance" (describing states with "independent judicial system[s] and legal framework[s]").[129]

This Part explores this question through archived material and a series of interviews with the policy executives charged with creating the moderation systems for YouTube and Facebook. It concludes that three

---

[124] Each of these platforms can of course be thought of differently. Facebook is primarily categorized as a social network site, *see* danah m. boyd & Nicole B. Ellison, *Social Network Sites: Definition, History, and Scholarship*, 13 J. COMPUTER-MEDIATED COMM. 210, 210 (2008); YouTube is seen as video-sharing; and Twitter is seen as both a social network and an RSS newsfeed. But all of these sites have one thing in common: they host, publish, and moderate user-generated content. This Article will look at these platforms in that capacity only.

[125] R. A. W. Rhodes, *The New Governance: Governing Without Government*, 44 POL. STUD. 652, 652 (1996). Indeed, the idea of Facebook as a nation-state or government, in the traditional sense, has been analyzed and dismissed. Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807, 1807 (2012) (concluding "regulatory power [over Facebook] is, de facto, dispersed across a wide array of international actors"). Professor Frank Pasquale has described these platforms as "feudal" or "sovereigns," FRANK PASQUALE, THE BLACK BOX SOCIETY 140–68, 187–218 (2015) (arguing that terms of service or contracts are inappropriate or ineffective remedies in an essentially "feudal" sphere, *id.* at 144, and arguing that platforms act as "sovereign[s]" over realms of life, *id.* at 163, 189), while Professor Rory Van Loo has called them "digital regulators," Van Loo, *supra* note 100, at 1267.

[126] Orly Lobel, *The Renew Deal: The Fall of Regulation and the Rise of Governance in Contemporary Legal Thought*, 89 MINN. L. REV. 342, 405 (2004).

[127] *Id.* at 406.

[128] *Id.*

[129] Adrian Leftwich, *Governance, Democracy and Development in the Third World*, 14 THIRD WORLD Q. 605, 610 (1993).

main factors influenced the development of these platforms' moderation systems: (1) an underlying belief in free speech norms; (2) a sense of corporate responsibility; and (3) the necessity of meeting users' norms for economic viability.

A. Platforms' Baseline in Free Speech

Conversations with the people who were in charge of creating the content-moderation regimes at these platforms reveal that they were indeed influenced by the concerns about user free speech and collateral censorship raised in *Zeran*.

1. Free Speech Norms. — For those closely following the development of online regulation, § 230 and *Zeran* were obvious foundational moments for internet speech. But at the time, many online commercial platforms did not think of themselves as related to speech at all. As a young First Amendment lawyer in the Bay Area, Nicole Wong was an active witness to the development of private internet companies' speech policies.[130] In the first few years of widespread internet use, Wong recalled that very few lawyers were focusing on the responsibilities that commercial online companies and platforms might have toward moderating speech.[131] But as most major print newspapers began posting content on websites between 1996 and 1998, the overlap between speech and the internet became more noticeable.[132] Likewise, just as more traditional publishing platforms for speech were finding their place on the internet, new internet companies were discovering that they were not just software companies, but that they were also publishing platforms.[133] At first, Wong's clients were experiencing speech as only a secondary effect of their primary business, as in the case of *Silicon Investor*, a day-trading site that was having issues with the content published on its message boards.[134] Others, like Yahoo, were actively recognizing that online speech was an intractable part of their business models.[135] Despite this reality, the transition to thinking of themselves as speech platforms was still slow. "They had just gone public," Wong said of her representation of early Yahoo. "They had only two lawyers in their legal department. . . . [N]either had any background in First Amendment law or content moderation or privacy. They were corporate

---

130 Telephone Interview with Nicole Wong, Former Emp., Google (Apr. 2, 2016).

131 *Id.*

132 David Shedden, *New Media Timeline (1996)*, POYNTER. (Dec. 16, 2004), http://www.poynter.org/2004/new-media-timeline-1996/28775/ [https://perma.cc/M37E-AJHE] (listing examples of new media sites that launched "on the Web" during 1996, including *The New York Times*, *Toronto Star*, *Chicago Tribune*, *Miami Herald*, and *Washington Post*).

133 Telephone Interview with Nicole Wong, *supra* note 130.

134 *Id.*

135 *Id.*

lawyers."[136] The problem identified by Wong was that these new internet corporations still thought of themselves as software companies — they did not think about "the lingering effects of speech as part of what they were doing."[137] In facing these new challenges, Wong had become one of the few people not only in Silicon Valley, but also in the United States, capable of advising on these challenges, with her background in First Amendment doctrine, communications, and electronic privacy.[138]

Wong's expertise led her to join Google full time in 2004. In October 2006, Google acquired YouTube, the popular online video site, and Wong was put in charge of creating and implementing content-moderation policies.[139] Creating the policies regarding what type of content would be acceptable on YouTube had an important free speech baseline: legal content would not be removed unless it violated site rules.[140] Wong and her content-moderation team actively worked to try to make sure these rules did not result in overcensorship of user speech. One such moment occurred in late December 2006, when two videos of Saddam Hussein's hanging surfaced on YouTube shortly after his death. One video contained grainy footage of the hanging itself; the other contained video of Hussein's corpse in the morgue. Both videos violated YouTube's community guidelines at the time — though for slightly different reasons. "The question was whether to keep either of them up," said Wong, "and we decided to keep the one of the hanging itself, because we felt from a historical perspective it had real value."[141] The second video was deemed "gratuitous violence" and removed from the site.[142] A similarly significant exception occurred in June 2009, when a video of a dying Iranian Green Movement protestor shot in the chest and bleeding from the eyes was ultimately kept on YouTube because of its political significance.[143] YouTube's policies and internal guidelines on violence were altered to allow for the exception.[144] In 2007, a video was uploaded to YouTube of a man being brutally beaten by four men in a cell and was removed for gratuitous violence in violation of

---

136 *Id.*

137 *Id.*

138 For an example of Wong's insight into these issues, see ELECTRONIC MEDIA AND PRIVACY LAW HANDBOOK (Nicole Wong et al. eds., 2003).

139 Telephone Interview with Nicole Wong, *supra* note 130; Rosen, *supra* note 12.

140 Site rules for impermissible content were related to banning content that was otherwise legal but that contained things like graphic violence or overt sexual activity. Buni & Chemaly, *supra* note 12; *see also infra* pp. 1632–33.

141 Telephone Interview with Nicole Wong, *supra* note 130.

142 *Id.*

143 Buni & Chemaly, *supra* note 12.

144 *Id.* It is important to make a distinction between "policies," which were the public rules posted for users about what content was allowed, and the internal "rules" that sites used to moderate speech. As will be shown in section III.A, *infra* pp. 1631–35, platforms' internal rules to moderate content came years before public policies were posted. The internal rules were also more detailed.

YouTube's community guidelines.[145] Shortly after, however, it was restored by Wong and her team after journalists and protestors contacted YouTube to explain that the video was posted by Egyptian human rights activist Wael Abbas to inform the international community of human rights violations by the police in Egypt.[146]

At Facebook, there was a similar slow move to organize platform policies on user speech. It was not until November 2009, five years after the site was founded, that Facebook created a team of about twelve people to specialize in content moderation.[147] Like YouTube, Facebook hired a lawyer, Jud Hoffman, to head their Online Operations team as Global Policy Manager. Hoffman recalled that, "when I got there, my role didn't exist."[148] Hoffman was charged with creating a group separate from operations that would formalize and consolidate an ad hoc draft of rules and ensure that Facebook was transparent with users by publishing a set of "Community Standards."[149] The team consisted of six people in addition to Hoffman, notably Dave Willner, who had created a first draft of these "all-encompassing" rules, which contained roughly 15,000 words.[150]

At Twitter, the company established an early policy not to police user content, except in certain circumstances, and rigorously defended that right.[151] Adherence to this ethos led to Twitter's early reputation among social media platforms as "the free speech wing of the free speech

---

[145] Neal Ungerleider, *Why This Ex-White House Tech Honcho Is Now Working on Human Rights*, FAST COMPANY (June 18, 2015), https://www.fastcompany.com/3046409/why-this-ex-white-house-tech-honcho-is-now-working-on-human-rights [https://perma.cc/52F4-JWD8].

[146] Telephone Interview with Nicole Wong, *supra* note 130.

[147] Telephone Interview with Dave Willner, Former Head of Content Policy, Facebook & Charlotte Willner, Former Safety Manager, User Operations, Facebook (Mar. 23, 2016).

[148] Telephone Interview with Jud Hoffman, Former Glob. Policy Manager, Facebook (Jan. 22, 2016).

[149] *Id.* "Community Standards" is Facebook's term for its *public* content-moderation policies. It is important to note that the internal rules created by Dave Willner predated the public Community Standards for the site. The internal rules informed, in part, the creation and substance of Facebook's public policies.

[150] *Id.*

[151] Sarah Jeong, *The History of Twitter's Rules*, MOTHERBOARD (Jan. 14, 2016, 10:00 AM), http://motherboard.vice.com/read/the-history-of-twitters-rules [https://perma.cc/X34U-HF4A]; *see also The Twitter Rules*, TWITTER SUPPORT (Jan. 18, 2009), https://web.archive.org/web/20090118211301/ttp://twitter.zendesk.com/forums/26257/entries/18311 [https://perma.cc/SMM6-NZEU]. Its rules' spartan nature was a purposeful reflection of the central principles and mission of the company. A preamble that accompanied the Twitter Rules from 2009 to 2016 reads:

> Our goal is to provide a service that allows you to discover and receive content from sources that interest you as well as to share your content with others. We respect the ownership of the content that users share and each user is responsible for the content he or she provides.

Id. "Because of these principles, we do not actively monitor user's content and will not censor user content, except in limited circumstances . . . ." *Id.*

party."[152] It also meant that unlike YouTube and Facebook, which actively took on content moderation of their users' content, Twitter developed no internal content-moderation process for taking down and reviewing content. The devotion to a fundamental free speech standard was reflected not only in what Twitter did not do to police user content, but also in what it did to protect it. Alexander Macgillivray joined Twitter as General Counsel in September 2009, a position he held for four years.[153] In that time, Macgillivray regularly resisted government requests for user information and user takedown. "We value the reputation we have for defending and respecting the user's voice," Macgillivray stated in 2012.[154] "We think it's important to our company and the way users think about whether to use Twitter, as compared to other services."[155]

A common theme exists in all three of these platforms' histories: American lawyers trained and acculturated in American free speech norms and First Amendment law oversaw the development of company content-moderation policy. Though they might not have "directly imported First Amendment doctrine," the normative background in free speech had a direct impact on how they structured their policies.[156] Wong, Hoffman, and Willner all described being acutely aware of their predisposition to American democratic culture, which put a large emphasis on free speech and American cultural norms. Simultaneously, there were complicated implications in trying to implement those American democratic cultural norms within a global company. "We were really conscious of not just wholesale adopting a kind of U.S. jurisprudence free expression approach," said Hoffman.[157] "[We would] try to step back and focus on the mission [of the company]."[158] Facebook's mission is to "[g]ive people the power to build community and bring the world closer together."[159] But even this, Willner acknowledged, is "not a cultural-neutral mission. . . . The idea that the world

---

[152] Josh Halliday, *Twitter's Tony Wang: "We Are the Free Speech Wing of the Free Speech Party*," THE GUARDIAN (Mar. 22, 2012, 11:57 AM), http://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/QR8B-CW74].

[153] Somini Sengupta, *Twitter's Free Speech Defender*, N.Y. TIMES (Sept. 2, 2012) [hereinafter Sengupta, *Twitter's Free Speech Defender*], http://nyti.ms/2GlWiKy [https://perma.cc/VM7K-99RJ]; Somini Sengupta, *Twitter General Counsel Leaves as Company Prepares to Go Public*, N.Y. TIMES: BITS (Aug. 30, 2013, 3:52 PM), https://bits.blogs.nytimes.com/2013/08/30/twitter-general-counsel-leaves-as-company-prepares-to-go-public/ [https://perma.cc/97RM-EB2W].

[154] Sengupta, *Twitter's Free Speech Defender*, *supra* note 153.

[155] *Id.*

[156] Telephone Interview with Jud Hoffman, *supra* note 148.

[157] *Id.*

[158] *Id.*

[159] *About*, FACEBOOK, https://www.facebook.com/pg/facebook/about/ [https://perma.cc/3ZV5-MECX].

should be more open and connected is not something that, for example, North Korea agrees with."[160]

2. Government Request and Collateral Censorship Concerns. — Beyond holding general beliefs in the right to users' free speech, these platforms have also implemented policies to protect user speech from the threat of government request and collateral censorship.[161]

Twitter's early pushback to government requests related to its users' content is well documented. In his time as General Counsel, Macgillivray regularly resisted government requests for user information and user takedown. In January 2011, he successfully resisted a federal gag order over a subpoena in a grand jury investigation into Wikileaks.[162] "[T]here's not yet a culture of companies standing up for users when governments and companies come knocking with subpoenas looking for user data or to unmask an anonymous commenter who says mean things about a company or the local sheriff," said *Wired* of Twitter's resistance to the gag order.[163] "Twitter deserves recognition for its principled upholding of the spirit of the First Amendment."[164] Despite the victory over the gag order, Twitter was eventually forced to turn over data to the Justice Department after exhausting all its appeals.[165] A similar scenario played out in New York, when a judge ordered Twitter to supply all the Twitter posts of Malcolm Harris, an Occupy Wall Street protester charged with disorderly conduct.[166] There, too, Twitter lost, but not before full resort to the appeals process.[167]

---

160 Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

161 This is not to say that collateral censorship issues are not a concern with private platforms' content-moderation systems. To the contrary, there are also many well-documented instances where platforms have cooperated with government requests for takedown and raised serious collateral censorship concerns. This section simply tries to give an overview of when platforms have proactively sought to avoid these concerns, even though doing so is costly and not necessary under § 230. *See* Balkin, *supra* note 11, at 2298–99 (explaining how the government can offer both carrots and sticks to entice private entities to cooperate with it regarding speech regulation); *see also, e.g.*, Emma Llansó, *German Proposal Threatens Censorship on Wide Array of Online Services*, CTR. FOR DEMOCRACY & TECH.: BLOG (Apr. 7, 2017), https://cdt.org/blog/german-proposal-threatens-censorship-on-wide-array-of-online-services/ [https://perma.cc/W9QT-5MP9] (discussing the dangers of allowing government units to flag issues for takedown using private content-moderation platforms).

162 Ryan Singel, *Twitter's Response to WikiLeaks Subpoena Should Be the Industry Standard*, WIRED (Jan. 10, 2011, 7:56 PM), https://www.wired.com/2011/01/twitter-2/ [https://perma.cc/5DV4-6JPN].

163 *Id.*

164 *Id.*

165 Sengupta, *Twitter's Free Speech Defender*, *supra* note 153.

166 *Id.*

167 Naomi Gilens, *Twitter Forced to Hand Over Occupy Wall Street Protester Info*, ACLU: FREE FUTURE (Sept. 14, 2012, 5:28 PM), https://www.aclu.org/blog/national-security/twitter-forced-hand-over-occupy-wall-street-protester-info [https://perma.cc/9UUT-F56Y].

Wong also described regularly fighting government requests to take down certain content, collateral censorship, and the problems with applying American free speech norms globally. For example, in November 2006, the Thai government announced that it would block YouTube to anyone using a Thai IP address unless Google removed twenty offensive videos from the site.[168] While some of the videos "clearly violated the YouTube terms of service," others simply featured Photoshopped images of the King of Thailand with feet on his head.[169] In Thailand, insulting the King was illegal and punishable by as much as fifteen years in prison.[170] Nicole Wong was hard pressed to find the content offensive. "My first instinct was it's a cartoon. It's a stupid Photoshop," she stated, "but then it suddenly became a kind of learning moment for me about international speech standards versus First Amendment speech standards and there was a lot more American First Amendment exceptionalism [in that space] than previously."[171] Wong traveled to Thailand to resolve the dispute and was overwhelmed by the popular love she observed in the Thai people for their King. "You can't even imagine [their love for their King]," she recounted of the trip:

> Every Monday literally eighty-five percent of the people show up to work in a gold or yellow shirt and dress[172] and there's a historical reason for it: the only source of stability in this country is this King . . . They absolutely revere their King. . . . Someone at the U.S. Embassy described him as a "blend of George Washington, Jesus, and Elvis." Some people . . . tears came to their eyes as they talked about the insults to the King and how much it offended them. That's the part that set me back. Who am I, a U.S. attorney sitting in California to tell them: "No, we're not taking that down. You're going to have to live with that."[173]

After the trip, Wong and her colleagues agreed to remove the videos within the geographical boundaries of Thailand, with the exception of critiques of the military.[174]

A few months later, events similar to those in Thailand emerged, but ended in a different result. In March 2007, Turkey blocked access to YouTube for all Turkish users in response to a judge-mandated order.[175] The judgment came in response to a parody news broadcast that jokingly quipped that the founder of modern Turkey, Mustafa Kemal

---

168 Rosen, *supra* note 12.

169 *Id.*

170 *Lese-Majeste Explained: How Thailand Forbids Insult of Its Royalty*, BBC NEWS (Oct. 6, 2017), http://www.bbc.com/news/world-asia-29628191 [https://perma.cc/58GZ-X7YZ].

171 Telephone Interview with Nicole Wong, *supra* note 130.

172 Yellow is the color associated with the King in Thailand. *Profile: Thailand's Reds and Yellows*, BBC NEWS (July 13, 2012), http://www.bbc.com/news/world-asia-pacific-13294268 [https://perma.cc/K79R-5AWP] (calling yellow "the king's colour").

173 Telephone Interview with Nicole Wong, *supra* note 130.

174 *Id.*

175 Rosen, *supra* note 12.

Atatürk, was gay.[176] As with the King in Thailand, ridicule or insult of Atatürk was illegal in Turkey. Though the video had already been voluntarily removed, Turkey had searched and provided Google with a list of dozens of similarly offensive videos and demanded their takedown.[177] Unwilling to meet the blanket demand, Wong and her colleagues at Google found themselves parsing the intricacies of Turkish law on defamation of Atatürk, measuring those standards against the videos highlighted as offensive by the Turkish government, and then offering compromises to ban in Turkey only those videos that they found actually violated Turkish law.[178] This seemed to strike an accord for a period of time.[179] A little over a year later, however, in June 2007, the Turkish government demanded Google ban access to all such videos not only in Turkey, but worldwide.[180] Google refused, and Turkey subsequently blocked YouTube throughout Turkey.[181]

All three platforms faced the issue of free speech concerns versus censorship directly through platform rules or collateral censorship by government request when a video called *Innocence of Muslims* was uploaded to YouTube.[182] Subtitled "The Real Life of Muhammad," the video depicts Muslims burning the homes of Egyptian Christians, before cutting to "cartoonish" images that paint Muhammad as a bastard, homosexual, womanizer, and violent bully.[183] The video's negative depiction of the Muslim faith sparked a firestorm of outrage in the Islamic world and fostered anti-Western sentiment.[184] As violence moved from Libya to Egypt, YouTube issued a statement that while the video would remain posted on the site because the content was "clearly within [its] guidelines," access to the video would be temporarily restricted in Libya and Egypt.[185]

At Facebook, the debate between violation of platform guidelines versus concerns over collateral censorship also played out. By the time the video was posted, many of Facebook's difficulties with hate speech had been distilled into a single rule: attacks on institutions (for example,

---

176 Jeffrey Rosen, *The Delete Squad*, NEW REPUBLIC (Apr. 29, 2013), https://newrepublic.com/article/113045/free-speech-internet-silicon-valley-making-rules [https://perma.cc/XB7Q-BSBA].

177 Rosen, *supra* note 12.

178 *Id.*

179 *Id.*

180 *Id.*

181 *Id.*

182 *The Anti-Islam-Film Riots: A Timeline*, THE WEEK (Sept. 18, 2012), http://theweek.com/articles/472285/antiislamfilm-riots-timeline [https://perma.cc/V6TK-N8M3].

183 David D. Kirkpatrick, *Anger over a Film Fuels Anti-American Attacks in Libya and Egypt*, N.Y. TIMES (Sept. 11, 2012), http://nyti.ms/2BxU77K [https://perma.cc/JZJ8-5QUD].

184 *Id.*

185 Eva Galperin, *YouTube Blocks Access to Controversial Video in Egypt and Libya*, ELECTRONIC FRONTIER FOUND.: DEEPLINKS BLOG (Sept. 12, 2012), https://www.eff.org/deeplinks/2012/09/youtube-blocks-access-controversial-video-egypt-and-libya [https://perma.cc/Y25N-PESU].

countries, religions, or leaders) would be considered permissible content and stay up, but attacks on groups (people of a certain religion, race, or country) would be taken down.[186] In application, this meant that statements like "I hate Islam" were permissible on Facebook, while "I hate Muslims" was not. Hoffman, Willner, and their team watched the video, found no violative statements against Muslims, and decided to keep it on the site.[187] A few weeks later, the Obama Administration called on YouTube to reconsider leaving the video up, in part to quell the violence abroad.[188] Both YouTube and Facebook stuck to their decisions.[189] Reviewing this moment in history, Professor Jeffrey Rosen spoke to the significance of their decisions for collateral censorship: "In this case . . . the mobs fell well outside of U.S. jurisdiction, and the link between the video and potential violence also wasn't clear. . . . Had YouTube made a different decision . . . millions of viewers across the globe [would have been denied] access to a newsworthy story and the chance to form their own opinions."[190]

The early history and personnel of these companies demonstrate how American free speech norms and concerns over censorship became instilled in the speech policies of these companies. But they also raise a new question: if all three companies had § 230 immunity and all valued their users' free speech rights, why did they bother curating at all?

B. Why Moderate At All?

These online platforms have broad freedom to shape online expression and a demonstrated interest in free speech values. So why do they bother to create intricate content-moderation systems to remove speech?[191] Why go to the trouble to take down and then reinstate videos of violence like those Wong described? Why not just keep them up in the first place? The answers to these questions lead to the incentives for platforms to minimize online obscenity put in place by the Good Samaritan provision of § 230. Platforms create rules and systems to curate speech out of a sense of corporate social responsibility, but also, more importantly, because their economic viability depends on meeting users' speech and community norms.

1. Corporate Responsibility and Identity. — Some platforms choose to moderate content that is obscene, violent, or hate speech out of a sense

---

186 Rosen, *supra* note 176.

187 *Id.*

188 Claire Cain Miller, *Google Has No Plans to Rethink Video Status*, N.Y. TIMES (Sept. 14, 2012), http://nyti.ms/2swdwTK [https://perma.cc/DKJ8-VBG4].

189 *Id.*

190 Rosen, *supra* note 176.

191 These systems are discussed in detail in Part III, *infra* pp. 1630–62.

of corporate responsibility.[192] At YouTube, Wong looked to the values of the company in addition to American free speech norms in developing an approach to content moderation.[193] "Not everyone has to be a free-wheeling, free speech platform that is the left wing of the left wing party," she said, referring to Twitter's unofficial content-moderation policy:

> But you get to decide what the tone and tenor of your platform look[] like, and that's a First Amendment right in and of itself. Yahoo or Google had a strong orientation toward free speech, [and] being more permissive of a wide range of ideas and the way those ideas are expressed, they created community guidelines to set what [users] can come here for, because they want the largest possible audience to join.[194]

Like Wong, Hoffman and Willner considered the mission of Facebook — "to make the world more open and connected"[195] — and found that it often aligned with larger American free speech and democratic values.[196] These philosophies were balanced against competing principles of user safety, harm to users, public relations concerns for Facebook, and the revenue implications of certain content for advertisers.[197] The balance often favored free speech ideals of "leaving content up" while at the same time trying to figure out new approaches or rules that would still satisfy concerned users and encourage them to connect and interact on the platform.[198] "We felt like Facebook was the most important platform for this kind of communication, and we felt like it was our responsibility to figure out an answer to this," said Hoffman.[199]

Likewise, Twitter's corporate philosophy of freedom of speech justified its failure to moderate content.[200] In recent years, Twitter's approach has started to change. In a *Washington Post* editorial, the new General Counsel of Twitter, Vijaya Gadde, used very different rhetoric

---

192 *See* Citron & Norton, *supra* note 104, at 1455 n.119 ("Such decisions may be justified as a matter of corporate law under the social entity theory of the corporation, which permits corporate decision-makers to consider and serve the interests of all the various constituencies affected by the corporation's operation." (citing Lisa M. Fairfax, *Doing Well While Doing Good: Reassessing the Scope of Directors' Fiduciary Obligations in For-Profit Corporations with Non-Shareholder Beneficiaries*, 59 WASH. & LEE L. REV. 409, 412 (2002))).

193 Telephone Interview with Nicole Wong, *supra* note 130.

194 *Id.*

195 *See Note from Mark Zuckerberg*, FACEBOOK (Apr. 27, 2016), https://newsroom.fb.com/news/2016/04/marknote/ [https://perma.cc/E7P5-SZZX]. Facebook changed its mission statement last year to "giv[ing] people the power to build community and bring[ing] the world closer together." Mark Zuckerberg, *Post*, FACEBOOK (June 22, 2017), https://www.facebook.com/zuck/posts/10154944663901634 [https://perma.cc/3PCE-KN9H]; *FAQs*, FACEBOOK: INV. REL., https://investor.fb.com/resources/default.aspx [https://perma.cc/AF3Q-WNFX].

196 *See* Telephone Interview with Jud Hoffman, *supra* note 148; *see also* Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

197 Telephone Interview with Jud Hoffman, *supra* note 148.

198 *Id.*

199 *Id.*

200 *See supra* pp. 1620–21.

than that of her predecessor: "Freedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up," wrote Gadde.[201] "We need to do a better job combating abuse without chilling or silencing speech."[202] Over the last two years, the company has slowly made good on its promise, putting a number of policies and tools in place to make it easier for users to filter and hide content they do not want to see.[203]

2. Economic Reasons. — Though corporate responsibility is a noble aim, the primary reason companies take down obscene and violent material is the threat that allowing such material poses to potential profits based in advertising revenue.[204] Platforms' "sense of the bottom-line benefits of addressing hate speech can be shaped by consumers' — i.e., users' — expectations."[205] If a platform creates a site that matches users' expectations, users will spend more time on the site and advertising revenue will increase.[206] Take down too much content and you lose not only the opportunity for interaction, but also the potential trust of users. Likewise, keeping up all content on a site risks making users uncomfortable and losing page views and revenue. According to Willner and Hoffman, this theory underlies much of the economic rationale behind Facebook's extensive moderation policies.[207] As Willner stated, "Facebook is profitable only because when you add up a lot of tiny interactions worth nothing, it is suddenly worth billions of dollars."[208] Wong spoke

---

[201] Vijaya Gadde, Editorial, *Twitter Executive: Here's How We're Trying to Stop Abuse While Preserving Free Speech*, WASH. POST (Apr. 16, 2015), http://wapo.st/1VyRio4 [https://perma.cc/G4BD-BLNC].

[202] *Id.*

[203] Kate Klonick, *Here's What It Would Take for Twitter to Get Serious About Its Harassment Problem*, VOX (Oct. 25, 2016, 10:50 AM), http://www.vox.com/new-money/2016/10/25/13386648/twitter-harassment-explained [https://perma.cc/VA7M-TRTH]. It is important to note that these methods used by Twitter to maximize free speech by shielding the viewer are really just a type of shadow censorship.

[204] *See* Citron & Norton, *supra* note 104, at 1454 n.113 ("[T]he traditional 'shareholder primacy' view . . . understands the corporation's primary (and perhaps exclusive) objective as maximizing shareholder wealth." (first citing Mark J. Roe, *The Shareholder Wealth Maximization Norm and Industrial Organization*, 149 U. PA. L. REV. 2063, 2065 (2001); then citing A. A. Berle, Jr., *For Whom Corporate Managers Are Trustees: A Note*, 45 HARV. L. REV. 1365, 1367–69 (1932))).

[205] *Id.*

[206] Paul Alan Levy, *Stanley Fish Leads the Charge Against Immunity for Internet Hosts — But Ignores the Costs*, PUB. CITIZEN: CONSUMER L. & POL'Y BLOG (Jan. 8, 2011), http://pubcit.typepad.com/clpblog/2011/01/stanley-fish-leads-the-charge-against-immunity-for-internet-hosts-but-ignores-the-costs.html [https://perma.cc/APS9-49BC] (arguing that websites that fail to provide protections against abuse will find "that the ordinary consumers whom they hope to serve will find it too uncomfortable to spend time on their sites, and their sites will lose social utility (and, perhaps more cynically, they know they will lose page views that help their ad revenue)"); *see also* Citron & Norton, *supra* note 104, at 1454 (discussing "digital hate as a potential threat to profits").

[207] Telephone Interview with Jud Hoffman, *supra* note 148; Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

[208] Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

of the challenge to meet users' expectations online slightly differently: as platforms attempting to catch up to changing social norms online.[209] Changing expectations about speech are happening both at the platform level, and also at a societal level, said Wong, who referred to the last twenty years of online speech as undergoing a "norm-setting process" that is developing at light speed in comparison to any other kind of publication platform.[210] "What we're still in the middle of is how do we think about . . . the norms of behavior when what's appropriate is constantly reiterated," said Wong.[211] "If you layer over all of that the technology change and the cultural, racial, national, [and] global perspectives, it's all just changing dramatically fast. It's enormously difficult to figure out those norms, let alone create policy to reflect them."[212] Nevertheless, reflecting these rapidly changing norms, and, accordingly, encouraging and facilitating platform interactions — users posting, commenting, liking, and sharing content — is how platforms like Facebook and YouTube have stayed in business and where platforms like Twitter have run into trouble.

Twitter's transformation from internet hero for its blanket refusal to police users' content to internet villain happened relatively swiftly. Though public awareness of online hate speech and harassment was already growing, the GamerGate controversy in 2014 raised new levels of global awareness about the issue.[213] As the least policed or rule-based platform, much of the blame fell on Twitter.[214] By 2015, the change in cultural values and expectations began to be reflected in new public standards and policy at Twitter. The site added new language prohibiting "promot[ing] violence against others . . . on the basis of race, ethnicity, national origin, religion, sexual orientation, gender, gender identity, age, or disability" to the Twitter Rules and prohibited revenge

---

209 Telephone Interview with Nicole Wong, *supra* note 130.

210 *Id.*

211 *Id.*

212 *Id.*

213 In August of that year, anonymous users targeted a number of women in the gaming industry — including game developers Zoë Quinn, Brianna Wu, and critic Anita Sarkeesian — in a series of harassment campaigns across multiple platforms, including Twitter. Jason Schreier, *Thousands Rally Online Against Gamergate*, KOTAKU (Oct. 15, 2014, 10:48 AM), https://kotaku.com/thousands-rally-online-against-gamergate-1646500492 [https://perma.cc/7E49-9KJA]. The harassment efforts included doxing, as well as rape and death threats. Sarah Kaplan, *With #GamerGate, the Video-Game Industry's Growing Pains Go Viral*, WASH. POST (Sept. 12, 2014), http://wapo.st/2EvoOJ7 [https://perma.cc/C4YH-B7J6]. The widespread and graphic nature of the controversy shifted norms and led to many calls on social media platforms to take a more proactive stance against online harassment and hate speech. Schreier, *supra*.

214 Charlie Warzel, *"A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment*, BUZZFEED: NEWS (Aug. 11, 2016, 9:43 AM), https://www.buzzfeed.com/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s [https://perma.cc/GQD5-XX97].

porn.[215] On December 30, 2015, Twitter published a new set of Twitter Rules — which were largely nothing new, but rather an official incorporation of the separate pages and policies in one place.[216] In January 2016, one Twitter spokesperson described the changes: "Over the last year, we have clarified and tightened our policies to reduce abuse, including prohibiting indirect threats and nonconsensual nude images. Striking the right balance will inevitably create tension, but user safety is critical to our mission at Twitter and our unwavering support for freedom of expression."[217]

In the mid-1990s, Post presciently wrote about how this interplay between users' norms around speech and content of online platforms would play out. Post suggested competition between individual online platforms would result in a "market for rules," which would allow users to seek networks that have "rule sets" to their liking.[218] At least with regard to Twitter, this platform-exit prediction is mostly accurate. Over the last few years, many users unhappy with the policies of Twitter left the platform and favored other platforms like Facebook, Instagram, and Snapchat.[219] As Twitter's user growth stagnated, many blamed the site's inability to police harassment, hate speech, and trolling on its site for the slump.[220] In late 2016, Twitter announced a host of new services for users to control their experience online, block hate speech and harassment, and control trolls.[221] Post's idea of a "market for rules" is an incredibly useful heuristic to understand the history of online content moderation, with two small updates: (1) the history of Twitter reveals a nuance not fully predicted by Post — that is, rather than exit a platform, some users would stay and expect platforms to alter rule sets and policies reactively in response to user pressure; and (2) the "market for rules"

---

215 Jeong, *supra* note 151 (alterations in original); *see also* Issie Lapowsky, *Why Twitter Is Finally Taking a Stand Against Trolls*, WIRED (Apr. 21, 2015, 2:14 PM), https://www.wired.com/2015/04/twitter-abuse/ [https://perma.cc/4V7R-43VK].

216 *See* @megancristina, *Fighting Abuse to Protect Freedom of Expression*, TWITTER: BLOG (Dec. 30, 2015), https://blog.twitter.com/official/en_au/a/2015/fighting-abuse-to-protect-freedom-of-expression-au.html [https://perma.cc/PP5E-KKAE].

217 Jeong, *supra* note 151.

218 Post, *supra* note 101, at para. 42.

219 *See* Jeff Dunn, *Here's How Slowly Twitter Has Grown Compared to Facebook, Instagram, and Snapchat*, BUS. INSIDER (Feb. 10, 2017, 6:14 PM), http://www.businessinsider.com/twitter-vs-facebook-snapchat-user-growth-chart-2017-2 [https://perma.cc/PF6U-GGPC] (assuming arguendo that growth in market alone cannot account for the slow growth of users on Twitter as compared to the growth of users on social media platforms like Snapchat, Instagram, and Facebook).

220 *See, e.g.*, Sarah Frier, *Twitter Fails to Grow Its Audience, Again*, BLOOMBERG TECH. (July 27, 2017, 7:00 AM), https://bloom.bg/2uFyz3B [https://perma.cc/XN9N-74BH]; Umair Haque, *The Reason Twitter's Losing Active Users*, HARV. BUS. REV. (Feb. 12, 2016), https://hbr.org/2016/02/the-reason-twitters-losing-active-users [https://perma.cc/KH4W-MK7P]; Joshua Topolsky, *The End of Twitter*, NEW YORKER (Jan. 29, 2016), https://www.newyorker.com/tech/elements/the-end-of-twitter [https://perma.cc/VZH3-V94L].

221 Klonick, *supra* note 203.

paradigm mistakes the commodity at stake in online platforms. The commodity is not just the user, but rather it is the content created and engaged with by a user culture.[222] In this sense there is no competition between social media platforms themselves, as Post suggests, because they are complementary, not substitute, goods.[223]

Whether rooted in corporate social responsibility or profits, the development of platforms' content-moderation systems to reflect the normative expectations of users is precisely what the creation of the Good Samaritan provision in § 230 sought. Moreover, the careful monitoring of these systems to ensure user speech is protected can be traced to the free speech concerns of § 230 outlined in *Zeran*. The answer to the dilemma of what § 230 protects — immunity for good actors creating decency online or protection against collateral censorship — seems not to be an either/or answer. Rather, both purposes seem to have an essential role to play in the balance of private moderation of online speech.

With this new knowledge about the motivations behind platforms' content-moderation systems, we can then ask the next question in the debate over internet intermediaries: how are platforms actually moderating? The answer to this question, explored in the next Part, is essential to understanding how platforms should — or should not — be understood for the purposes of First Amendment law.

III. How Are Platforms Governing? The Rules, Process, and Revision of Content-Moderation Systems

Much of the analysis over how to categorize online platforms with respect to the First Amendment is missing a hard look at what these platforms are actually doing and how they are doing it. In part, this is because the private content-moderation systems of major platforms like Facebook, Twitter, and YouTube are historically opaque. This Part seeks to demonstrate how these systems actually work to moderate online speech. In doing this, Part III looks at the history of how content-moderation systems changed from those of standards to those of rules, how platforms enforce these rules, and how these rules are subject to change. Many of these features bear remarkable resemblance to heuristics and structures familiar in legal decisionmaking. Despite these similarities, platform features are best thought of not in terms of First

---

[222] *See* Balkin, *supra* note 7, at 4–6.

[223] Moreover, Post's free-market idea of user exit is also challenged by current studies. In an ongoing project, the Electronic Frontier Foundation has worked to document and present evidence of the negative psychological impact that leaving — either by choice or by banning — certain social media platforms can have on users. *See Submit Report*, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/submit-report [https://perma.cc/25NK-LGA2] (offering a platform for users to report erroneous or unjust account deactivations). These studies support the theory Lessig describes in *Code: Version 2.0*, in which he proffers that leaving an internet platform is more difficult and costly than expected. *See* LESSIG, *supra* note 21, at 288–90.

Amendment doctrine — as reflecting the role of a state actor, a broadcaster, or a newspaper editor — but in terms of a private self-regulatory system to govern online speech.

A. Development of Moderation: From Standards to Rules

When Dave Willner joined a small team to specialize in content moderation in November 2009, no public "Community Standards" existed at Facebook. Instead, all content moderation was based on one page of internal "rules" applied globally to all users. Willner recalled that the moderation policies and guidance for enforcing them were limited.[224] "The [policy] guidance was about a page; a list of things you should delete: so it was things like Hitler and naked people. None of those things were wrong, but there was no explicit framework for why those things were on the list."[225] Willner's now-wife Charlotte was also working at Facebook doing customer service and content moderation and had been there for a year before Dave joined.[226] She described the ethos of the pre-2008 moderation guidelines as "if it makes you feel bad in your gut, then go ahead and take it down."[227] She recalled that the "Feel bad? Take it down" rule was the bulk of her moderation training prior to the formation of Dave's group in late 2008.[228] Wong described a similar ethos in the early days at YouTube, especially around efforts to know when to remove graphic violence from the site. Speaking of reinstating the 2007 video of the Egyptian protestor being brutally beaten,[229] Wong said: "It had no title on it. It wasn't posted by him. . . . I had no way of knowing what it was and I had taken something down that had real significance as a human rights document. So we put it back up. And then we had to create another exception to the no-violence rule."[230]

Though both Wong and the Willners used the term "rule" in describing these prescriptions for takedown, a more precise term for these early guidelines might be "standard." In legal theory, the "rules-standards conflict" describes the battle between two formal resolutions for legal controversy.[231] An example of a standard is "don't drive too fast." An

---

[224] Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

[225] *Id.*

[226] *Id.*

[227] *Id.*

[228] *Id.*

[229] *See supra* pp. 1619–20.

[230] Telephone Interview with Nicole Wong, *supra* note 130.

[231] *See generally* MARK KELMAN, A GUIDE TO CRITICAL LEGAL STUDIES 40–45 (1987); Pierre Schlag, *Rules and Standards*, 33 UCLA L. REV. 379, 381–83 (1985); Anthony J. Casey & Anthony Niblett, *The Death of Rules and Standards* 7–10 (Univ. of Chi. Coase-Sandor Inst. for Law & Econ., Paper No. 738, 2015), https://ssrn.com/abstract=2693826 [https://perma.cc/4WJV-43RM]; Lawrence Solum, *Legal Theory Lexicon: Rules, Standards, and Principles*, LEGAL THEORY BLOG (Sept. 6, 2009, 9:40 AM), http://lsolum.typepad.com/legaltheory/2009/09/legal-theory-lexicon-rules-standards-and-principles.html [https://perma.cc/XR4C-9QT3].

example of a rule is a speed limit set at sixty-five miles per hour. There are trade-offs to picking one as the formal solution over the other. Standards are often "restatements of purpose" or values,[232] but because they are often vague and open ended, they can be "subject to arbitrary and/or prejudiced enforcement" by decisionmakers.[233] This purposive approach, however, can also mean that standards are enforced precisely and efficiently and can be more accommodating to changing circumstances. Rules, on the other hand, have the issues reverse to those of standards. Rules are comparatively cheap and easy to enforce, but they can be over- and underinclusive and, thus, can lead to unfair results.[234] Rules permit little discretion and in this sense limit the whims of decisionmakers, but they also can contain gaps and conflicts, creating complexity and litigation.[235]

Whichever approach is used, a central point is that the principles formalized in rules and standards are rooted in the social norms and values of a community.[236] Standards are more direct analogues of values or purpose but "require[] that the enforcing community . . . come to some consensus on the meaning of a value term."[237] Rules are more distant from the norms they are based on and "do not depend on ongoing dialogue to gain dimension or content . . . even by someone who shares no sense of community with his fellows."[238]

The development at YouTube and Facebook from standards to rules for content moderation reflects these trade-offs. A simple standard against something like gratuitous violence is able to reach a more tailored and precise measure of justice that reflects the norms of the community, but it is vague, capricious, fact dependent, and costly to enforce.

This can be seen at YouTube, which in mid-2006 employed just sixty workers to review all video that had been flagged by users for all reasons.[239] For violations of terms of service, one team of ten, deemed the Safety, Quality, and User Advocacy Department, or SQUAD, worked in shifts "around the clock" to keep YouTube from "becoming a shock site."[240] That team was given a one-page bullet-point list of standards that instructed on removal of things like animal abuse, videos showing blood, visible nudity, and pornography.[241] A few months later, in the

---

232 KELMAN, *supra* note 231, at 40 (emphasis omitted).
233 *Id.* at 41.
234 *Id.* at 40.
235 *See id.* at 40–47.
236 *See* Eric A. Posner, *Standards, Rules, and Social Norms*, 21 HARV. J.L. & PUB. POL'Y 101, 107–11 (1997).
237 KELMAN, *supra* note 231, at 61.
238 *Id.* at 62.
239 Buni & Chemaly, *supra* note 12.
240 *Id.*
241 *Id.*

fall of 2006, the YouTube list turned into a six-page booklet drafted with input from the SQUAD, Wong, and other YouTube lawyers and policy executives.[242] Five years later, in 2011, the volume of uploaded video to YouTube had more than doubled in size, making delicate, precise decisions less feasible.[243] In addition, the content-moderation team had expanded and been outsourced. Accordingly, the more individually tailored standards against gratuitous violence had slowly been replaced by precise rules, which were easier and less costly to enforce. Moderators were given a booklet with internal rules for content moderation. This booklet was regularly annotated and republished with changes to moderation policies and rules.[244] Many of these new rules were drafted as "exceptions" to rules. Eventually, a more detailed iterative list of rules and their exceptions largely replaced the standards-based approach of earlier years.

Similar to the experience at YouTube, Facebook eventually abandoned the standards-based approach as the volume of user-generated content increased, the user base diversified, and the content moderators globalized. Dave Willner was at the helm of this transition. Though Facebook had been open globally for years, Willner described much of the user base during his early days there as still relatively homogenous — "mostly American college students" — but that was rapidly changing as mobile technology improved and international access to the site grew.[245] Continuing to do content moderation from a single list of banned content seemed untenable and unwieldy. Instead, Willner set about changing the entire approach:

> In the early drafts we had a lot of policies that were like: "Take down all the bad things. Take down things that are mean, or racist, or bullying." Those are all important concepts, but they're value judgments. You have to be more granular and less abstract than that. Because if you say to forty college students [content moderators], "delete all racist speech," they are not going to agree with each other about what's racist.[246]

Eliminating standards that evoked nonobservable values, feelings, or other subjective reactions was central to Willner's new rulebook for moderation. Instead, he focused on the implicit logic of the existing page of internal guidelines and his experience and extrapolated from them to

---

[242] *Id.*

[243] On May 1, 2009, YouTube had twenty hours of video upload per minute; by May 1, 2011, forty-eight hours of video were uploaded per minute. *See* Mark R. Robertson, *500 Hours of Video Uploaded to YouTube Every Minute [Forecast]*, TUBULAR INSIGHTS (Nov. 13, 2015), http://tubularinsights.com/hours-minute-uploaded-youtube/ [https://perma.cc/A9Q7-N3VM].

[244] Buni & Chemaly, *supra* note 12.

[245] Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147; *see also* Buni & Chemaly, *supra* note 12.

[246] Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

create objective rules.[247] The first draft of these "all-encompassing" rules was written largely by Willner in 2009 and contained roughly 15,000 words.[248] The end goal was consistency and uniformity: to get the same judgment on a piece of content, regardless of who was moderating it.[249]

Exactly "who" was moderating the content changed significantly in January 2009, when Facebook opened its office in Dublin and first started outsourcing its content moderation through consulting groups. Before then, most moderators worked in Palo Alto and were similar to Facebook's main user base — "homogenous college students."[250] The shift to outsourced moderation continued when a new community operations team was set up in Hyderabad, India.[251] Around the same time, Hoffman joined Facebook's team as Global Policy Manager with the goal of formalizing and consolidating the rules Willner had started to draft, and ensuring that Facebook was transparent with users by publishing a set of public rules in the form of "Community Standards."[252]

Hoffman and Willner worked together to transform the early ad hoc abuse standards into operational internal rules for content moderators, a document that today is over eighty pages long.[253] This movement from standards to rules was "ultimately a form of technical writing," said Willner.[254] "You cannot tell people to delete photos with ugly clothes in them. You have to say 'delete photos with orange hats in them.'"[255] For Willner, some of the hardest parts of defining categories, elements, and distinctions came in moderating art and nudity.[256] For Hoffman, it was more difficult to create rules around hate speech. "We couldn't make a policy that said 'no use of the N-word at all,'" he recalled, describing the difficulty in policing racial slurs.[257] "That could be completely insensitive to the African American community in the United States. But you also don't want it used as hate speech. So it's

---

247 *Id.*

248 *Id.*

249 *Id.*

250 *Id.*

251 *Id.*

252 *Id.* "Community Standards" is Facebook's term for its *public* content-moderation policies. It is important to note that the internal rules created by Willner predated the public Community Standards for the site. In fact, it was the internal rules that informed, in part, the creation and substance of Facebook's public policies.

253 *Id.*

254 *Id.*

255 *Id.*

256 "Art doesn't exist as a property of an image. There are no art pixels that you can find in images we think are classy or beautiful or uplifting. . . . But what we realized about art was that [moderation] questions about art weren't about art itself, it was about art being an exception to an existing restriction. . . . [S]o the vast majority of art is fine. It's when you're talking about things that might meet the definition of nudity or racism or violence that people think are important." *Id.*

257 Telephone Interview with Jud Hoffman, *supra* note 148.

almost impossible to turn that into an objective decision because context matters so much."[258] The answer was to turn context into a set of objective rules. In evaluating whether speech was likely to provoke violence, for example, Hoffman and his team developed a four-part test to assess credible threats: time, place, method, and target.[259] If a post specified any three of these factors, the content would be removed, and if appropriate, authorities notified.[260]

Content moderation at YouTube and Facebook developed from an early system of standards to an intricate system of rules due to (1) the rapid increase in both users and volume of content; (2) the globalization and diversity of the online community; and (3) the increased reliance on teams of human moderators with diverse backgrounds. The next section discusses enforcement of these rules.

B. How the Rules Are Enforced: Trained Human Decisionmaking

Content moderation happens at many levels. It can happen before content is actually published on the site, as with ex ante moderation, or after content is published, as with ex post moderation. These methods can be either *reactive*, in which moderators passively assess content and update software only after others bring the content to their attention, or *proactive*, in which teams of moderators actively seek out published content for removal. Additionally, these decisions can be *automatically* made by software or *manually* made by humans.[261] The majority of

---

[258] *Id.*

[259] Univ. of Hous. Law Ctr., *UH Law Center and the ADL Present Racists, Bigots and the Law on the Internet*, YOUTUBE (Oct. 10, 2012), https://youtu.be/aqqvYPyr6cI?list=UU3rht1s60KV8PnW1ds47_KQ [https://perma.cc/Q7SF-TYNM] (recording of Jud Hoffman, Glob. Policy Manager, Facebook).

[260] *Id.* Many situations, however, were lacking in context. Online bullying was the type of issue that often arose with insufficient background. As Hoffman described:

> There is a traditional definition of bullying — a difference in social power between two people, a history of contact — there are elements. But when you get a report of bullying, you just don't know. You have no access to those things. So you have to decide whether you're going to assume the existence of some of those things or assume away the existence of some of those things. Ultimately what we generally decided on was, "if you tell us that this is about you and you don't like it, and you're a private individual not a public figure, we'll take it down." Because we can't know whether all these other things happened, and we still have to make those calls. But I'm positive that people were using that function to game the system. . . . I just don't know if we made the right call or the wrong call or at what time.

Telephone Interview with Jud Hoffman, *supra* note 148. Hoffman's description also demonstrates two major drawbacks to using rules rather than standards. A blanket rule against bullying can simultaneously result in people manipulating a rule to "walk the line" and also result in permissible content being mistakenly removed. *Id.*

[261] *See* James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42, 63–70 (2015) (describing how moderation systems operate differently along several lines — automatic or manual, transparent or secret, ex ante or ex post, and centralized or decentralized). Professor James Grimmelmann's taxonomy, while foundational, speaks more generally to all of internet moderation

this section focuses on ex post *reactive* content moderation, specifically looking at the implementation of rules with respect to human decisionmaking, pattern recognition, and professionalization of judgment.

1. Ex Ante Content Moderation. — When a user uploads a video to Facebook, a message appears: "Processing Videos: The video in your post is being processed. We'll send you a notification when it's done and your post is ready to view."[262] Ex ante content moderation is the process that happens in this moment between "upload" and publication.[263] The vast majority of this moderation is an automatic process run largely through algorithmic screening without the active use of human decisionmaking.

An example of content that can be moderated by these methods is child pornography, which can reliably be identified upon upload through a picture-recognition algorithm called PhotoDNA.[264] Under federal law, production, distribution, reception, and possession of an image of child pornography is illegal, and as such, sites are obligated to remove it.[265] A known universe of child pornography — around 720,000 illegal images — exists online.[266] By converting each of these images to grayscale, overlaying a grid, and assigning a numerical value to each

---

rather than content-publishing platforms specifically. In the context of speech, the distinction between ex ante and ex post is especially important, in that it determines whether moderation is happening before or after publication. Of secondary concern is whether content is being moderated through reactive or proactive measures. Finally, the ultimate means of reaching decisions, whether through software or humans, is descriptively helpful, but less legally significant.

[262] *Videos*, FACEBOOK: HELP CTR., https://www.facebook.com/help/1542711413755 95/ [https://perma.cc/FHD2-4RAY].

[263] Because ex ante content moderation happens before publication takes place, it is the type of prior restraint that scholars like Balkin are concerned with. *See generally* Balkin, *supra* note 11. Of the two automatic means of reviewing and censoring content — algorithm and geoblocking — geoblocking is of more concern for the purposes of collateral censorship and prior restraint. In contrast, algorithms are currently used to remove illegal content like child pornography or copyright violations. *But see* Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986, 1003–05 (2008) (noting that the Digital Millennium Copyright Act's notice-and-takedown provisions give platforms no incentive to investigate and therefore "suppress critical speech as well as copyright infringement," *id.* at 1003).

[264] Tracy Ith, *Microsoft's PhotoDNA: Protecting Children and Businesses in the Cloud*, MICROSOFT: NEWS (July 15, 2015), https://news.microsoft.com/features/microsofts-photodna-protecting-children-and-businesses-in-the-cloud/ [https://perma.cc/H7F7-KSB7].

[265] *See* 18 U.S.C. §§ 2251–2252A (2012). It is important to remember that § 230 expressly states that no internet entity has immunity from federal criminal law, intellectual property law, or communications privacy law. 47 U.S.C. § 230(e) (2012). This means that every internet service provider, search engine, social networking platform, and website is subject to thousands of laws, including child pornography laws, obscenity laws, stalking laws, and copyright laws. *Id.*

[266] This "known universe" of child pornography is maintained and updated by the International Centre for Missing and Exploited Children and the U.S. Department of Homeland Security in a program known as Project Vic. Mark Ward, *Cloud-Based Archive Tool to Help Catch Child Abusers*, BBC NEWS (Mar. 24, 2014), http://www.bbc.com/news/technology-26612059 [https:// perma.cc/KX6E-C5R6].

square, researchers were able to create a "hash," or signature, that remained even if the images were altered.[267] As a result, platforms can determine whether an image contains child pornography in the microseconds between upload and publication.[268] Geoblocking is another form of automatic ex ante moderation. Unlike PhotoDNA, which prevents the publication of illegal content, geoblocking prevents both publication and viewing of certain content based on a user's location. As happened in the controversy over the *Innocence of Muslims* video, geoblocking usually comes at the request of a government notifying a platform that a certain type of posted content violates its local laws.[269]

Of course, algorithms do not decide for themselves which kind of content they should block from being posted. Content screened automatically is typically content that can reliably be identified by software and is illegal or otherwise prohibited on the platform. This universe of content that is automatically moderated ex ante is regularly evaluated and updated through iterative software updates and machine learning. For example, in a similar fashion to PhotoDNA, potential copyright violations can be moderated proactively through software like Content ID. Developed by YouTube, Content ID allows creators to give their content a "digital fingerprint" so it can be compared against other uploaded content.[270] Copyright holders can also flag already-published copyright violations through notice and takedown.[271] These two systems work together, with user-flagged copyrighted material eventually added to ContentID databases for future proactive review.[272] This mix of proactive, manual moderation and informed, automatic ex ante moderation is also evident in the control of spam. All three platforms (and most internet companies, generally) struggle to control spam postings on their sites. Today, spam is mostly blocked automatically from publication through software. Facebook, Twitter, and YouTube, however, all feature mechanisms for users to report spam manually.[273] Ex ante screening software is iteratively updated to reflect these flagged spam sources.

---

[267] *Id.*

[268] Ith, *supra* note 264.

[269] *See supra* pp. 1624–25; *see also, e.g.*, Telephone Interview with Nicole Wong, *supra* note 130.

[270] *How Content ID Works*, YOUTUBE: HELP, https://support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/RZ5T-9UPN].

[271] *See, e.g.*, *Submit a Copyright Takedown Notice*, YOUTUBE: HELP, https://support.google.com/youtube/answer/2807622 [https://perma.cc/DAS6-8G3R].

[272] *How Content ID Works*, *supra* note 270.

[273] *See, e.g.*, *How Twitter Aims to Prevent Your Timeline from Filling Up with Spam*, PANDA MEDIACENTER (Sept. 12, 2014), http://www.pandasecurity.com/mediacenter/social-media/twitter-spam/ [https://perma.cc/8HM8-G63Z]; James Parsons, *Facebook's War Continues Against Fake Profiles and Bots*, HUFFINGTON POST (Mar. 22, 2015, 5:03 PM), http://www.huffingtonpost.com/james-parsons/facebooks-war-continues-against-fake-profiles-and-bots_b_6914282.html [https://perma.cc/3X6E-7AJ7].

2. Ex Post Proactive Manual Content Moderation. — Recently, a form of content moderation that harkens to the earlier era of AOL chat rooms has reemerged: platforms proactively seeking out and removing published content. Currently, this method is largely confined to the moderation of extremist and terrorist speech. As of February 2016, dedicated teams at Facebook have proactively removed all posts or profiles with links to terrorist activity.[274] Such efforts were doubled in the wake of terrorist attacks.[275] This is an important new development affecting content moderation, which seeks to strike an ever-evolving balance between competing interests: ensuring national security and maintaining individual liberty and freedom of expression. While a topic worthy of deep discussion, it is not the focus of this paper.[276]

3. Ex Post Reactive Manual Content Moderation. — With the exception of proactive moderation for terrorism described above, almost all user-generated content that is published is reviewed *reactively*, that is, through ex post flagging by other users and review by human content moderators against internal guidelines. Flagging — alternatively called reporting — is the mechanism provided by platforms to allow users to express concerns about potentially offensive content.[277] The adoption by social media platforms of a flagging system serves two main functions: (1) it is a "practical" means of reviewing huge volumes of content, and (2) its reliance on users serves to legitimize the system when platforms are questioned for censoring or banning content.[278]

Facebook users flag over one million pieces of content worldwide every day.[279] Content can be flagged for a variety of reasons, and the vast majority of items flagged do not violate the Community Standards of Facebook. Instead content flags often reflect internal group conflicts or disagreements of opinion.[280] To resolve the issue, Facebook created a new reporting "flow" — the industry term to describe the sequence of screens users experience as they make selections — that encourages users to resolve issues themselves rather than report them for review to

---

[274] Natalie Andrews & Deepa Seetharaman, *Facebook Steps Up Efforts Against Terrorism*, WALL ST. J. (Feb. 11, 2016, 7:39 PM), http://on.wsj.com/1TVJNse [https://perma.cc/9CY7-BYD9]. As will be discussed later, corporate censorship of speech at the behest or encouragement of governments raises questions of collateral censorship and state action doctrine. *See infra* pp. 1658–62.

[275] Andrews & Seetharaman, *supra* note 274.

[276] For an excellent, thorough, and cutting-edge discussion of this issue, see Danielle Keats Citron, *Extremist Speech and Compelled Conformity*, 93 NOTRE DAME L. REV. (forthcoming 2018), https://ssrn.com/abstract=2941880 [https://perma.cc/6WM2-H8PY].

[277] Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 NEW MEDIA & SOC'Y 410, 411 (2016).

[278] *Id.* at 412.

[279] *See* Buni & Chemaly, *supra* note 12; Telephone Interview with Monika Bickert, Head of Glob. Policy Mgmt., Facebook & Peter Stern, Head of Policy Risk Team, Facebook (Jan. 19, 2016).

[280] *The Trust Engineers*, RADIOLAB (Feb. 9, 2015, 8:01 PM), https://www.radiolab.org/story/trust-engineers/ [https://perma.cc/9C4N-SJDW].

Facebook.[281] Users reporting content first click a button to "Report/Mark as Spam," which then quickly guides users to describe their report in terms like "Hate Speech," "Violence or Harmful Behavior," or "I Don't Like This Post."[282] Some types of reports, such as harassment or self-harm, guide users to the option of "social reporting" — a tool that "enables people to report problematic content not only to Facebook, but also directly to their friends to help resolve conflicts."[283] To enhance the response time of content moderation, the reporting flow also has the instrumental purpose of triaging flagged content for review.[284] This makes it possible for Facebook to immediately prioritize certain content for review and, when necessary, notify authorities of emergency situations like suicide, imminent threats of violence, terrorism, or self-harm. Other content, like possible hate speech, nudity, pornography, or harassment, can be queued into less urgent databases for general review.[285]

After content has been flagged to a platform for review, the precise mechanics of the decisionmaking process become murky. The "army" of content moderators and "[t]he details of moderation practices are routinely hidden from public view," write Catherine Buni and Soraya Chemaly.[286] "[S]ocial media companies do not publish details of their internal content moderation guidelines; no major platform has made such guidelines public."[287] These internal guidelines also change much more frequently than the public Terms of Service or Community Standards. Focusing largely on Facebook, except where specified, the next section seeks to illuminate this process by integrating previously published information together with interviews of content moderators and platform internal guidelines. The system of people making the decisions will be examined first followed by a review of the internal guidelines that inform that decisionmaking process.

(a) Who Enforces the Rules? — When content is flagged or reported, it is sent to a server where it awaits review by a human content moderator.[288] At Facebook, there are three basic tiers of content moderators: "Tier 3" moderators, who do the majority of the day-to-day reviewing

---

[281] *Id.*

[282] Alexei Oreskovic, *Facebook Reporting Guide Shows How Site Is Policed (Infographic)*, HUFFINGTON POST (June 19, 2012, 9:38 PM), https://www.huffingtonpost.com/2012/06/20/facebook-reporting-guide_n_1610917.html [https://perma.cc/4LHU-BLGD] .

[283] *Id.*

[284] Univ. of Hous. Law Ctr., *supra* note 259 (Jud Hoffman speaking).

[285] *Id.*

[286] Buni & Chemaly, *supra* note 12.

[287] *Id.*

[288] Skype Interviews with Kumar S. (Jan. 29–Mar. 9, 2016); Skype Interviews with Selahattin T. (Mar. 2–Mar. 11, 2016); Skype Interview with Jagruti (Jan. 26, 2016) [hereinafter Content Moderator Interviews]. These content moderators were Tier 3 workers based in India and Eastern Europe and provided background on what the process looks like from the perspective of a content moderator.

of content; "Tier 2" moderators, who supervise Tier 3 moderators and review prioritized or escalated content; and "Tier 1" moderators, who are typically lawyers or policymakers based at company headquarters.[289]

In the early days, recent college graduates based in the San Francisco Bay Area did much of the Tier 3 content moderation.[290] Today, most platforms, including Facebook, either directly employ content-moderation teams or outsource much of their content-moderation work to companies like oDesk (now Upwork), Sutherland, and Deloitte.[291] In 2009, Facebook opened an office in Dublin, Ireland, that had twenty dedicated support and user-operations staff.[292] In 2010, working with an outsourcing partner, Facebook opened a new office in Hyderabad, India, for user support.[293]

Today, Tier 3 moderators typically work in "call centers"[294] in the Philippines, Ireland, Mexico, Turkey, India, or Eastern Europe.[295] Within Facebook, these workers are called "community support" or "user support teams."[296] When working, moderators will log on to computers and access the server where flagged content is awaiting review.[297] Tier 3 moderators typically review material that has been flagged as a lower priority by the reporting flow. At Facebook, for example, this includes, in part, reports of nudity or pornography, insults or attacks based on religion, ethnicity, or sexual orientation, inappropriate or annoying content, content that is humiliating, or content that advocates violence to a person or animal.[298]

Tier 2 moderators are typically supervisors of Tier 3 moderators or specialized moderators with experience judging content. They work both remotely (many live in the United States and supervise groups that

---

289 Telephone Interview with J.L., Tier 2 Moderator, Facebook (Mar. 11, 2016). J.L. was a Tier 2 moderator based in the Eastern United States.

290 Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147; Telephone Interview with Sasha Rosse, Manager of Glob. Outsourcing, Facebook (May 16, 2016); Buni & Chemaly, *supra* note 12.

291 Telephone Interview with Sasha Rosse, *supra* note 290; Adrian Chen, *Inside Facebook's Outsourced Anti-Porn and Gore Brigade, Where "Camel Toes" Are More Offensive than "Crushed Heads*," GAWKER (Feb. 16, 2012, 3:45 PM), http://gawker.com/5885714/inside-facebooks-outsourced-anti-porn-and-gore-brigade-where-camel-toes-are-more-offensive-than-crushed-heads [https://perma.cc/HU7H-972C]; Chen, *supra* note 12.

292 Telephone Interview with Sasha Rosse, *supra* note 290.

293 *Id.*

294 Buni & Chemaly, *supra* note 12.

295 Content Moderator Interviews, *supra* note 288; Telephone Interview with Sasha Rosse, *supra* note 290; Chen, *supra* note 291; Chen, *supra* note 12.

296 Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147; Telephone Interview with Jud Hoffman, *supra* note 148.

297 Content Moderator Interviews, *supra* note 288; Telephone Interview with Sasha Rosse, *supra* note 290.

298 Telephone Interview with J.L., *supra* note 289.

are internationally based) and locally at call centers.[299] Tier 2 moderators review content that has been prioritized, like imminent threats of violence, self-harm, terrorism, or suicide. This content comes to Tier 2 directly through the reporting flow or by being identified and escalated to Tier 2 by Tier 3 moderators. Tier 2 moderators also review certain randomized samples of Tier 3 moderation decisions. In order to ensure the accuracy of moderation, Facebook and other platforms have a certain amount of built-in redundancy: the same piece of content is often given to multiple Tier 3 workers. If the judgment on the content varies, the content is reassessed by a Tier 2 moderator.[300]

Tier 1 moderation is predominantly performed at the legal or policy headquarters of a platform. At Facebook, for example, a Tier 3 worker could be based in Hyderabad, a Tier 2 supervisor could be based in Hyderabad, or remotely in a place like Dublin, but a Tier 1 contact would be based in Austin, Texas, or the San Francisco Bay Area. "There were not many levels between the boots-on-ground moderator and Menlo Park," stated one former Tier 2 supervisor who had worked at Facebook until 2012, speaking on the condition of anonymity.[301] "If I had doubts on something, I'd just send it up the chain."[302]

Recently, issues of scaling this model have led platforms to try new approaches to who enforces the rules. At YouTube, a new initiative was launched in late 2016 called the Heroes program, which deputizes users to actively participate in the content-moderation process in exchange for perks such as "access to exclusive workshops and sneak preview product launches."[303] Similarly, after a video of the murder of an elderly man in Cleveland stayed up for over an hour on Facebook, Zuckerberg announced the company would hire 3000 additional content moderators, increasing the size of the content-moderation team by two-thirds.[304]

(b) How Are the Rules Enforced? — As previously discussed, the external policy — or Community Standards — provided to the public is not the same as the internal rulebook used by moderators when trying to assess whether content violates a platform's terms of service. An analysis of the internal guidelines reveals a structure that in many ways

---

[299] *Id.*; Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

[300] Telephone Interview with J.L., *supra* note 289.

[301] *Id.*

[302] *Id.*

[303] Sarah Perez, *YouTube Enlists Volunteers to Moderate Its Site via a New "YouTube Heroes" Program*, TECHCRUNCH (Sept. 21, 2016), https://techcrunch.com/2016/09/21/youtube-enlists-volunteers-to-moderate-its-site-via-a-new-youtube-heroes-program/ [https://perma.cc/5HJN-K8E2]. *See generally Get Involved with YouTube Contributors*, YOUTUBE: HELP, https://support.google.com/youtube/answer/7124236 [https://perma.cc/6G72-ZUFT].

[304] Alex Heath, *Facebook Will Hire 3000 More Moderators to Keep Deaths and Crimes from Being Streamed*, BUS. INSIDER (May 3, 2017, 10:35 AM), http://www.businessinsider.com/facebook-to-hire-3000-moderators-to-keep-suicides-from-being-streamed-2017-5 [https://perma.cc/5LCM-QG59].

replicates the decisionmaking process present in modern jurisprudence. Content moderators act in a capacity very similar to that of a judge: moderators are trained to exercise professional judgment concerning the application of a platform's internal rules and, in applying these rules, moderators are expected to use legal concepts like relevance, reason through example and analogy, and apply multifactor tests.

(i) Training. — Willner and Hoffman's development of objective internal rules at Facebook was a project that became an essential element in the shift to content-moderation outsourcing made in early 2010.[305] While Facebook's Community Standards were applied globally, without differentiation along cultural or national boundaries,[306] content moderators, in contrast, came with their own cultural inclinations and biases. In order to ensure that the Community Standards were enforced uniformly, it was necessary to minimize content moderators' application of their own cultural values and norms when reviewing content and instead impose Facebook's.[307] The key to all of this was providing intensive in-person training on applying the internal rules. "It all comes down to training," stated Sasha Rosse, who worked with Willner to train the first team in Hyderabad:

> I liked to say that our goal was [to have a training system and rules set] so I could go into the deepest of the Amazon, but if I had developed parameters that were clear enough I could teach someone that had no exposure to anything outside of their village how to do this job.[308]

---

[305] Facebook outsourced only a small subset of reports in 2010. Most of the content-moderation work was still being performed by full-time employees and contractors in Hyderabad, Austin, and Palo Alto. Email from Jud Hoffman, Former Glob. Policy Manager, Facebook (Aug. 18, 2016) (on file with author).

[306] It is worth noting why Facebook has this policy. According to Willner, in writing the internal rules and the Community Standards, Facebook:

> realized that the nature of the product made regional rules untenable. There are no "places" in Facebook — there are just people with different nationalities, all interacting in many shared forums. Regional rules would make cross-border interactions and communities largely incoherent and moderation very hard if not impossible. For example, if a Greek user insults Atatürk and a Turkish user reports it, whose rules apply?

Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

[307] Though often referred to as "neutral," Facebook's values and norms — and the rules that attempted to reflect them — were distinctly American. *See supra* section II.A, pp. 1618–25.

[308] Telephone Interview with Sasha Rosse, *supra* note 290. Despite the internal rules and training, cultural biases still crept into moderation, especially when judging subjective content. For example, in 2010 and 2011, the Facebook content-policy team was still struggling to refine its guidelines as it simultaneously began to train moderators in India. Rules on nudity were relatively clear-cut because nudity could in large part be reduced to observable characteristics that were either present or not in content. But harder questions arose regarding the Facebook rules banning certain kinds of sexualized content: a person could be entirely clothed, but in a highly sexual position. At some point in the training process in India, a group of workers were given a list of observable rules about a picture that made it impermissibly sexual, but at the bottom of the rules there was a more general "Feel Bad" standard: if you feel like something is otherwise sexual or pornographic, take it

Training moderators to overcome cultural biases or emotional reactions in the application of rules to facts can be analogized to training lawyers or judges. In the law, training lawyers and judges through law school and practice bestows a "specialized form of cognitive perception — what Karl Llewellyn called 'situation sense' — that reliably focuses their attention on the features of a case pertinent to its valid resolution."[309] Professor Dan Kahan calls this "professional judgment," but it might also be called "pattern recognition" after Professor Howard Margolis's study of expert and lay assessments of risk,[310] or even likened to the rapid, instinctual categorization used by chicken sexers, expert workers whose entire job is to determine the sex of baby chickens a day or two after the chickens hatch.[311] Regardless of the label, training content moderators involves a repetitive process to "override" cultural or emotional reactions and replace them with rational "valid" resolutions.[312]

Recent studies show that professionalized judgment can thwart cognitive biases, in addition to increasing attention to relevant information and reliable application of rules.[313] In a series of experiments, Kahan asked judges, lawyers, and law students with various political inclinations to assess legal problems that were "designed to trigger unconscious political bias in members of the general public."[314] Despite the presence of irrelevant but polarizing facts, judges, and to a lesser degree, lawyers, were largely in agreement in deciding legal cases presented to them in the study.[315] In contrast, law students and members of the general public reliably made decisions in keeping with their personal political views when presented with politically polarizing information.[316] Replication of the study expanded these findings beyond mere political ideologies to more general "cultural cognition," that is, the "unconscious influence of

down. That standard when applied to global content by a small group of moderators was predictably overrestrictive. "Within a day or two, we saw a spike of incorrect decisions," said Hoffman, "where people on this team in India were removing flagged content that portrayed open-mouth kissing." Telephone Interview with Jud Hoffman, *supra* note 148.

309 Dan M. Kahan et al., *"Ideology" or "Situation Sense"? An Experimental Investigation of Motivated Reasoning and Professional Judgment*, 164 U. PA. L. REV. 349, 354–55 (2016) (quoting KARL N. LLEWELLYN, THE COMMON LAW TRADITION: DECIDING APPEALS 59–61, 121–57, 206–08 (1960)).

310 *See id.* at 372 (citing HOWARD MARGOLIS, DEALING WITH RISK: WHY THE PUBLIC AND THE EXPERTS DISAGREE ON ENVIRONMENTAL ISSUES (1996)).

311 *See* RICHARD HORSEY, THE ART OF CHICKEN SEXING (2002), http://cogprints.org/3255/1/chicken.pdf [https://perma.cc/K8J4-GJSV].

312 Kahan et al., *supra* note 309, at 372.

313 *Id.* at 374. *See generally* Dan M. Kahan et al., *"They Saw a Protest": Cognitive Illiberalism and the Speech-Conduct Distinction*, 64 STAN. L. REV. 851 (2012) (explaining how cultural cognition shapes interpretations of legally relevant facts).

314 Kahan et al., *supra* note 309, at 354.

315 *Id.*

316 *Id.*

individuals' group commitments on their perceptions of legally consequential facts."[317]

The experiments by Kahan and his co-authors demonstrate empirically what Facebook learned through experience: people can be trained in domain-specific areas to overcome their cultural biases and to apply rules neutrally. Just as this truth is an essential part of the legal system, it is an essential part of Facebook's moderation system.

(ii) Similarities to American Law and Legal Reasoning. — Before applying law to facts, a judge must first determine which facts are relevant. Procedural rules like the Federal Rules of Evidence acknowledge that the inclusion of certain information may unfairly exploit decisionmakers' biases and emotions and, thus, provide guidance on how to exclude information from review.[318] At Facebook, the internal rules used by content moderators, or "Abuse Standards," similarly contain extensive guidance on what "relevant" content a moderator should review in assessing a report.[319]

Once a moderator has followed the procedural rules to narrow the relevant content to be reviewed, the actual Abuse Standards — or rules — must be applied. These start with a list of per se bans on content.[320] In Abuse Standards 6.2, these per se bans on content are lists of rules split into nine somewhat overlapping categories.[321] But as is typical of a rules-based approach, these lists contain as many exceptions as they do rules. In "Graphic Content," listed violations include any "[p]oaching of animals" as well as "[p]hotos and digital images showing internal organs, bone, muscle, tendons, etc.," while "[c]rushed heads,

---

317 Kahan et al., *supra* note 313, at 851.

318 *See* Kahan et al., *supra* note 309, at 365.

319 ODESK, ABUSE STANDARDS 6.1, https://www.scribd.com/doc/81863464/oDeskStandards [https://perma.cc/P6ZV-V9ZA] [hereinafter AS 6.1]; ODESK, ABUSE STANDARDS 6.2, https://www.scribd.com/doc/81877124/Abuse-Standards-6-2-Operation-Manual [https://perma.cc/2JQF-AWMY] [hereinafter AS 6.2]. These are copies of documents that were leaked from a content moderator working at oDesk (now Upwork) doing content moderation for Facebook. They are not the actual rules of Facebook, but they are oDesk's approximation of Facebook's rules. Charles Arthur, *Facebook's Nudity and Violence Guidelines Are Laid Bare*, THE GUARDIAN (Feb. 21, 2012, 4:36 PM), https://www.theguardian.com/technology/2012/feb/21/facebook-nudity-violence-censorship-guidelines [https://perma.cc/9LNL-L4C6]. For a more current but very similar version of these policies as expressed through content-moderator training documents, see Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence*, THE GUARDIAN (May 21, 2017, 1:00 PM), https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence [https://perma.cc/U7DY-5VHE].

320 AS 6.1, *supra* note 319; AS 6.2, *supra* note 319.

321 Those categories are "Sex and Nudity," "Illegal Drug Use," "Theft Vandalism and Fraud," "Hate Content," "Graphic Content," "IP Blocks and International Compliance," "Self Harm," "Bullying and Harassment," and "Credible Threats." AS 6.2, *supra* note 319, at 4. Among the twelve items under "Sex and Nudity," for example, are "[a]ny OBVIOUS sexual activity, even if naked parts are hidden from view by hands, clothes or other objects. Cartoons/art included. Foreplay allowed ([k]issing, groping, etc.) even for same-sex individuals" and "[p]eople 'using the bathroom.'" *Id.*

limbs, etc. are ok as long as no insides are showing."[322] Likewise, "mere depiction" of some types of content — "hate symbols" like swastikas, or depictions of Hitler or Bin Laden — are automatic violations, "unless the caption (or other relevant content) suggests that the user is not promoting, encouraging or glorifying the [symbol]."[323]

Some more complicated types of speech borrow from American jurisprudence for the structure of their rules. Under "Hate Content," a chart provides examples of "Protected Categories" and counsels moderators to mark "content that degrades individuals based on the . . . protected categories" as a violation.[324] A second chart on the page demonstrates how the identification of the type of person — ordinary persons, public figures, law enforcement officers, and heads of state — as well as their membership in a protected group will factor into the permissibility of the content.[325] All credible threats are to be escalated regardless of the "type of person."[326] These examples demonstrate the influence of American jurisprudence on the development of these rules. Reference to "Protected Categories" is similar to the protected classes of the Civil Rights Act of 1964.[327] The distinction between public and private figures is reminiscent of First Amendment, defamation, and invasion of privacy law.[328] The emphasis on credibility of threats harkens to the balance between free speech and criminal law.[329]

Beyond borrowing from the law substantively, the Abuse Standards borrow from the way the law is applied, providing examples and analogies to help moderators apply the rules. Analogical legal reasoning, the method whereby judges reach decisions by reasoning through analogy

---

[322] *Id.*

[323] *Id.* at 8.

[324] *Id.* at 5 (including race, ethnicity, national origin, religion, sex, gender identity, sexual orientation, disability, and serious disease as protected categories).

[325] *Id.* An empty threat against a public figure like Paul McCartney is permissible, but an empty threat against a head of state like President Barack Obama should be removed. Any type of content about a law enforcement officer — empty threat, credible threat, negative reference, cyberbullying, and attacks with hate symbols — is a violation under the Abuse Standards, as is any kind of attack based on being a victim of sexual assault. *Id.*

[326] "For safety and legal reasons, we consider threats credible if they:

1. Target heads of state or specific law enforcement officers . . . [;]
2. Contain 3/4 details: time, place, method, specific target (not impossible to carry out)[;]
3. Target people with a history of assassination attempt/s[;]
4. Include non-governmental bounties (promising earthly and heavenly rewards for a target's death)[.]"

Id. at 7.

[327] Pub. L. No. 88-352, §§ 201–202, 703, 78 Stat. 241, 243–44, 255–57 (outlawing discrimination based on race, color, religion, sex, or national origin).

[328] *See* RESTATEMENT (SECOND) OF TORTS § 558 (AM. LAW INST. 1977); *see also* Gertz v. Robert Welch, Inc., 418 U.S. 323, 351 (1974) (refusing to extend the *N.Y. Times Co. v. Sullivan*, 376 U.S. 254 (1964), standard for public officials' defamation claims to private individuals).

[329] *See, e.g.*, Brett A. Sokolow et al., *The Intersection of Free Speech and Harassment Rules*, 38 HUM. RTS. 19, 19 (2011).

between cases, is a foundation of legal theory.[330] Though the use of example and analogy plays a central role throughout the Abuse Standards,[331] the combination of legal rule and example in content moderation seems to contain elements of both rule-based legal reasoning and analogical legal reasoning. For example, after stating the rules for assessing credibility, the Abuse Standards give a series of examples of instances that establish credible or noncredible threats.[332] "I'm going to stab (method) Lisa H. (target) at the frat party (place)," states Abuse Standards 6.2, demonstrating a type of credible threat that should be escalated.[333] "I'm going to blow up the planet on new year's eve this year" is given as an example of a noncredible threat.[334] Thus, content moderators are not expected to reason directly from prior content decisions as in common law — but the public policies, internal rules, examples, and analogies they are given in their rulebook are informed by past assessments.

In many ways, platforms' evolution from "gut check" standards to more specific rules tracks the evolution of the Supreme Court's doctrine defining obscenity. In *Jacobellis v. Ohio*,[335] Justice Stewart wrote that he could not "intelligibly" define what qualified something as obscene, but famously remarked, "I know it when I see it."[336] Both Charlotte Willner, at Facebook, and Nicole Wong, at Google, described a similar intuitive ethos for removing material in the early days of the platforms' content-moderation policies.[337] Eventually, Facebook's and YouTube's moderation standards moved from these standards to rules. Likewise, over a series of decisions, the Court attempted to make the criteria for obscenity more specific — in *Miller v. California*,[338] the Court issued a

---

330 *See* LLOYD L. WEINREB, LEGAL REASON: THE USE OF ANALOGY IN LEGAL ARGUMENT (2005); Edward H. Levi, *An Introduction to Legal Reasoning*, 15 U. CHI. L. REV. 501 (1948). Among philosophers and legal theorists an important distinction can be made between "pure" analogical legal reasoning, which looks exclusively to the similarities and differences between cases without use of legal rules, and "pure" rule-based legal reasoning, which deduces exclusively from rules without case comparison. *See generally* LARRY ALEXANDER & EMILY SHERWIN, DEMYSTIFYING LEGAL REASONING 64–103 (2008); Cass R. Sunstein, Commentary, *On Analogical Reasoning*, 106 HARV. L. REV. 741 (1993). The Abuse Standards do not clearly point toward a "pure" version of either of these reasoning approaches.

331 This is especially true in the case of introducing new rules or policies to moderators. For example, Abuse Standards 6.2 introduces a "fresh policy" on sexually explicit language and sexual solicitation, and lists thirteen examples of content that should be removed or kept up under the policy. AS 6.2, *supra* note 319, at 6. In Abuse Standards 6.1, an entire page is devoted to samples of pictures that fall in or out of the various bans on sex and nudity, cartoon bestiality, graphic violence, animal abuse, or Photoshopped images. AS 6.1, *supra* note 319, at 4.

332 AS 6.2, *supra* note 319.

333 *Id.* at 7.

334 *Id.*

335 378 U.S. 184 (1964).

336 *Id.* at 197 (Stewart, J., concurring).

337 Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147; Telephone Interview with Nicole Wong, *supra* note 130.

338 413 U.S. 15 (1973).

three-part test to evaluate whether state statutes designed to regulate obscene materials were sufficiently limited.[339] None of the tests created by the Court, however, comes close to the specificity of the facts and exceptions used by platforms today.

To summarize, knowledge about the training of content moderators and Abuse Standards 6.1 and 6.2 tells us much about how the rules are enforced in content-moderation decisions. Content moderators act in a capacity very similar to that of judges: (1) like judges, moderators are trained to exercise professional judgment concerning the application of a platform's internal rules; and (2) in applying these rules, moderators are expected to use legal concepts like relevancy, reason through example and analogy, and apply multifactor tests.

4. Decisions, Escalations, and Appeals. — At Facebook, Tier 3 moderators have three decisionmaking options regarding content: they can "confirm" that the content violates the Community Standards and remove it, "unconfirm" that the content violates Community Standards and leave it up, or escalate review of the content to a Tier 2 moderator or supervisor.[340] The Abuse Standards describe certain types of content requiring mandatory escalations, such as: child nudity or pornography, bestiality, credible threats, self-harm, poaching of endangered animals, Holocaust denial, all attacks on Atatürk, maps of Kurdistan, and burning of Turkish Flags.[341] If a moderator has decided to ban content, a Facebook user's content is taken down, and she is automatically signed off of Facebook. When the user next attempts to sign in, she will be given the following message: "We removed the post below because it doesn't follow the Facebook Community Standards."[342] When she clicks "Continue," the user is told: "Please Review the Community Standards: We created the Facebook Community Standards to help make Facebook a safe place for people to connect with the world around them. Please read the Facebook Community Standards to learn what kinds of posts are allowed on Facebook."[343] The user then clicks "Okay" and is allowed to log back in. At Facebook, users who repeatedly have content removed face a gradual intensification of punishment: two removed posts in a certain amount of time, for example, might mean your account is suspended for twenty-four hours. Further violations of community standards can result in total bans. At YouTube, moderators had

---

339 *Id.* at 24.

340 AS 6.1, *supra* note 319; AS 6.2, *supra* note 319.

341 AS 6.2, *supra* note 319, at 4.

342 Screenshot of Facebook Removal Notice, ME.ME (June 13, 2017, 5:38 AM), https://me.me/i/29-21-01-am-facebook-we-removed-something-you-posted-we-15347765 [https://perma.cc/2BHA-936B].

343 Screenshot of Facebook Community Standards Notice, ME.ME (May 1, 2017, 1:09 PM), https://me.me/i/7-20-am-pao-94-facebook-please-review-the-community-standards-13463100 [https://perma.cc/LNQ9-HHRV].

a slightly different set of options for each piece of content: "Approve" let a video remain; "Racy" gave the video an 18+ year-old rating; "Reject" allowed a video to be removed without penalizing the poster; and finally, "Strike" would remove the video and issue a penalty to the poster's account.[344]

The ability of an individual user to appeal a decision on content takedown, account suspension, or account deletion varies widely between the three major platforms. Facebook allows an appeal of the removal of only a profile or page — not individual posts or content.[345] To initiate an appeal process, a user's account must have been suspended.[346] Appeals are reviewed by the Community Operations teams on a rolling basis and sent to special reviewers.[347] In contrast, at YouTube, account suspensions, "strikes" on an account, and content removal are all appealable.[348] Video strikes can be appealed only once, and if a decision to strike is upheld, there is a sixty-day moratorium on the appeal of any additional strikes.[349] An appeal also lies if an account is terminated for repeated violations.[350] At Twitter, any form of action related to the Twitter Rules can be appealed.[351] Users follow instructions on the app itself or provided in an email sent to notify users that content has been taken down.[352] Twitter also includes an intermediary level of removal called a "media filter" on content that might be sensitive.[353] Rather than totally remove the content, the platform requires users to click through a warning in order to see the content.[354] Appeals are handled by support teams that, when possible, will use specialized team members to review culturally specific content.[355]

C. System Revision and the Pluralistic System of Influence

Facebook's Abuse Standards do more than shed light on substantive rules on speech or the mechanisms behind its decisionmaking. They also demonstrate that the internal rules of content moderation are iteratively revised on an ongoing basis, and much more frequently than the external public-facing policy. This can be seen on the first page of Abuse

---

344 Buni & Chemaly, *supra* note 12.

345 *How to Appeal*, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/resources/how-to-appeal [https://perma.cc/YM9B-Q2KF].

346 *Id.*

347 Telephone Interview with J.L., *supra* note 289.

348 *How to Appeal*, *supra* note 345.

349 *Id.*

350 *Id.*

351 *Id.*

352 *Id.*

353 *Id.*

354 *Id.*

355 *Id.*

Standards 6.1, titled "Major changes since A[buse] S[tandards] 6.0," which contains a bulleted list of amendments and alterations to the previous set of rules.[356] The list is divided into groupings of roughly related policy changes.[357] "Added sexual language and solicitation policy," states the first bullet.[358] Halfway down the page after "Sex & Nudity issues clarified" is a bulleted section beginning "Graphic violence policies updated as follows."[359] In Abuse Standards 6.2, there are fewer updates, summarized broadly under one bullet point, "Policy Changes":

> Graphic Content with respect to animal insides
>
> Threshold and considerations for credible threats
>
> Caricatures of protected categories
>
> Depicting bodily fluids
>
> Screenshots or other content revealing personal information
>
> PKK versus Kurdistan flags
>
> Updated policy on photo-shopped images[360]

The differences between these two versions demonstrate that internal policies and the rules that reflect them are constantly being updated. This is because Facebook is attempting, in large part, to rapidly reflect the norms and expectations of its users.

But how are platforms made aware of these "dramatically fast"[361] changing global norms such that they are able to alter the rules? This section discusses four major ways platforms' content-moderation policies are subject to outside influence: (1) government request, (2) media coverage, (3) third-party civil society groups, and (4) individual users' use of the moderation process.

This multi-input content-moderation system is a type of pluralistic system.[362] Under the ideal theory, a pluralistic system consists of many

---

356 AS 6.1, *supra* note 319, at 2.

357 *Id.*

358 *Id.*

359 *Id.* "No exceptions for news or awareness-related context for graphic image depictions [—] confirm all such content; [h]uman/animal abuse subject to clear involvement/enjoyment/approval/encouragement by the poster [should be confirmed]; [e]ven fake/digital images of graphic content should be confirmed, but hand-drawn/cartoon/art images are ok." *Id.* (emphasis omitted).

360 AS 6.2, *supra* note 319, at 2.

361 Telephone Interview with Nicole Wong, *supra* note 130.

362 *See* Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, U.C. DAVIS L. REV. (forthcoming 2018), https://ssrn.com/abstract=3038939 [https://perma.cc/9HJS-NUZT]; *cf.* ROBERT A. DAHL, WHO GOVERNS? 197–99 (1961); DAVID HELD, MODELS OF DEMOCRACY 57–64 (2006); Freeman, *supra* note 15, at 559–60 ("In a pluralist 'interest representation' model of administrative law, administrative procedures and judicial review facilitate an essentially political decision-making process: They ensure that interest groups enjoy a forum in which to press their views and that agencies adequately consider those views when making policy choices.").

diverse external factions of equal strength competing to influence a neutral government.[363] In a perfect world, the competition between these minority factional interests serves to maintain equilibrium and representation of ideas in a democratic society.[364] But in practice, pluralism can be far from ideal.[365] This section discusses these interests and their potential democratic conflicts in the context of outside influence on platforms' content-moderation policies.

1. Government Requests.[366] — Lessig describes the architecture of the internet — or constitution — as built by an "invisible hand, pushed by government and by commerce."[367] Lessig does not describe these two forces as separate, but rather tandem in their effect. Thus far, this Article has principally focused on the commercial side of this dynamism, but platform architecture has also been informed by and subject to government interference. This interference can be through the more direct need to comply with local laws and jurisdictions, or by the more subtle influences of government lobbying and requests.

The previous examples of the Thai King, Atatürk, and *Innocence of Muslims*[368] illustrate how platforms have either conformed their policies, modified their policies, or rejected policy changes following government request. At YouTube, material would be removed within a country only if it violated the laws of that country — whether or not it was a violation was determined by YouTube's own lawyers.[369] If content was found to be in violation of a country's laws, a new policy would be issued and geoblocks put in place to prevent access to that content

---

[363] HELD, *supra* note 362, at 57–64.

[364] Freeman, *supra* note 15, at 560 ("Although conscious of capture, the theory envisions this pathology as limited to agencies, and as correctable, presumably by democratizing the agency decision-making process to include numerous interest groups. In this sense, interest representation reveals a lingering optimism about the democratic potential of pluralism, when properly structured."). *But see* Richard B. Stewart, *The Reformation of American Administrative Law*, 88 HARV. L. REV. 1667, 1713 (1975) (discussing how some interests like those of a regulated party may be overrepresented in agency government).

[365] *See* Freeman, *supra* note 15, at 560 (discussing the threat capture poses to democratic ideals of pluralism); *see also* Margot E. Kaminski, *When the Default Is No Penalty: Negotiating Privacy at the NTIA*, 93 DENV. L. REV. 925 (2016) (examining instances in which openness to participation by interest groups did not result in meaningful participation); David Thaw, *Enlightened Regulatory Capture*, 89 WASH. L. REV. 329 (2014) (examining instances in which regulatory capture by a concentrated interest group can be beneficial).

[366] For an excellent and more thorough discussion of the potential effects of government requests on online platforms and free speech, see Llansó, *supra* note 161.

[367] LESSIG, *supra* note 21, at 4.

[368] *See* section II.A.1–2, *supra* pp. 1618–25 (detailing how Wong established geoblocking within Thailand for some types of content — determined by YouTube — that ridiculed the King; how Wong established geoblocking within Turkey for some types of content — determined by YouTube — that disparaged Atatürk; and how Facebook and YouTube refused requests of the government to remove *Innocence of Muslims*, and instead kept it up as permissive under their own moderation rules and standards).

[369] Telephone Interview with Nicole Wong, *supra* note 130.

within that country.[370] Similar agreements were reached regarding depictions of Atatürk in Turkey.[371] At Facebook, however, content is not geoblocked but removed globally if international compliance requires.[372] Examples of this include support of the Kurdistan Workers' Party (PKK) or any content supporting Abdullah Ocalan.[373] Other types of content with specific geographic sensitivities, like Holocaust denial focusing on hate speech, attacks on Atatürk, maps of Kurdistan, and burning of Turkish flags, are required to be escalated.[374]

Twitter maintains a policy that it will take down posts only on request and only if they violate a country's laws.[375] This policy has occasionally been at odds with Twitter's more unofficial tactic of vigorously and litigiously protecting free speech.[376] Compromises have been reached, however, without sacrificing one for the other: in 2012, when India demanded Twitter remove a number of accounts that were fueling religious dissent, the company removed roughly half of the problematic accounts, but did so on the grounds that they violated Twitter's own policies for impersonation.[377] In other contexts, such as the requests for takedown in Egypt and Turkey, particularly during periods of revolution, Twitter has refused to capitulate to any government requests, and governments have consequently blocked the platform.[378]

Recently, however, platforms have been criticized for increasingly acquiescing to government requests, especially in the distribution of user information to police.[379] Platforms have also begun cooperating more proactively in response to the increased use of social media by the Islamic State of Iraq and Syria (ISIS) to recruit members and encourage terrorism. Over the last few years, all three sites have agreed to general requests from the United States and the United Nations to remove content related to ISIS or terrorism.[380] As discussed briefly in section III.B,

---

370 *Id.*

371 *Id.*

372 Telephone Interview with Dave Willner & Charlotte Willner, *supra* note 147.

373 AS 6.2, *supra* note 319, at 4.

374 *Id.*

375 Sengupta, *Twitter's Free Speech Defender*, *supra* note 153.

376 *Id.*

377 *Id.*

378 *See id.*; Sebnem Arsu, *Turkish Officials Block Twitter in Leak Inquiry*, N.Y. TIMES (Mar. 20, 2014), http://nyti.ms/2CpSGoI [https://perma.cc/RJN5-ULWT]; Erick Schonfeld, *Twitter Is Blocked in Egypt Amidst Rising Protests*, TECHCRUNCH (Jan. 25, 2011), https://techcrunch.com/2011/01/25/twitter-blocked-egypt/ [https://perma.cc/P8YD-QQ86].

379 *See, e.g.*, John Herrman, *Here's How Facebook Gives You Up to the Police*, BUZZFEED: NEWS (Apr. 6, 2012, 5:08 PM), https://www.buzzfeed.com/jwherrman/how-cops-see-your-facebook-account [https://perma.cc/GDK6-KPLT]; Dave Maass & Dia Kayyali, *Cops Need to Obey Facebook's Rules*, ELECTRONIC FRONTIER FOUND.: DEEPLINKS BLOG (Oct. 24, 2014), https://www.eff.org/deeplinks/2014/10/cops-need-obey-facebooks-rules [https://perma.cc/2RTA-8YZS].

380 *See* Andrews & Seetharaman, *supra* note 274; Joseph Menn & Dustin Volz, *Google, Facebook Quietly Move Toward Automatic Blocking of Extremist Videos*, REUTERS (June 24, 2016, 8:26 PM),

Facebook now maintains a team that is focused on terrorism-related content and helps promote "counter speech" against such groups.[381] The team actively polices terrorist pages and friend networks on the site. No posts from known terrorists are allowed on the site, even if the posts have nothing to do with terrorism. "If it's the leader of Boko Haram and he wants to post pictures of his two-year-old and some kittens, that would not be allowed," said Monika Bickert, Facebook's head of global policy management.[382] As Facebook has become more adept at and committed to removing such terrorism-related content, that content has moved to less restrictive platforms like Twitter. In just a four-month period in 2014, ISIS supporters used an estimated 46,000 Twitter accounts, though not all were active simultaneously.[383] Just before the dissemination of pictures of American journalist James Foley's beheading, the platform in 2015 began taking a different approach.[384] In early 2016, Twitter reported that it had suspended 125,000 accounts related to ISIS.[385]

2. Media Coverage. — The media do not have a major role in changing platform policy per se, but when media coverage is coupled with either (1) the collective action of users or (2) a public figure's involvement, platforms have historically been responsive.

An early high-profile example of media catalyzing collective action occurred around a clash between Facebook's nudity policy and breastfeeding photos posted by users. As early as 2008, Facebook received criticism for removing posts that depicted a woman breastfeeding.[386] The specifics of what triggered removal changed over time.[387] The changes came, in part, after a campaign in the media and in pages on Facebook itself staged partly by women who had their content removed.[388] Similar policy changes occurred after public outcry over

---

https://www.reuters.com/article/us-internet-extremism-video-exclusive/exclusive-google-facebook-quietly-move-toward-automatic-blocking-of-extremist-videos-idUSKCN0ZB00M [https://perma.cc/G3AT-J5K6].

381 Andrews & Seetharaman, *supra* note 274.

382 Telephone Interview with Monika Bickert & Peter Stern, *supra* note 279.

383 Julia Greenberg, *Why Facebook and Twitter Can't Just Wipe Out ISIS Online*, WIRED (Nov. 21, 2015, 7:00 AM), http://www.wired.com/2015/11/facebook-and-twitter-face-tough-choices-as-isis-exploits-social-media/ [https://perma.cc/QFU9-2N5V].

384 J.M. Berger, *The Evolution of Terrorist Propaganda: The Paris Attack and Social Media*, BROOKINGS INST.: TESTIMONY (Jan. 27, 2015), http://www.brookings.edu/research/testimony/2015/01/27-terrorist-propaganda-social-media-berger [https://perma.cc/28QK-HUJU].

385 Andrews & Seetharaman, *supra* note 274.

386 Mark Sweney, *Mums Furious as Facebook Removes Breastfeeding Photos*, THE GUARDIAN (Dec. 30, 2008, 8:17 AM), https://www.theguardian.com/media/2008/dec/30/facebook-breastfeeding-ban [https://perma.cc/Y3V4-X4EG].

387 *See* Soraya Chemaly, *#FreeTheNipple: Facebook Changes Breastfeeding Mothers Photo Policy*, HUFFINGTON POST (June 9, 2014, 6:48 PM), http://www.huffingtonpost.com/soraya-chemaly/freethenipple-facebook-changes_b_5473467.html [https://perma.cc/8TPP-JEGT].

388 *Id.*

Facebook's "real name" policy,[389] removal of a gay kiss,[390] censoring of an 1886 painting that depicted a nude woman,[391] posting of a beheading video,[392] and takedown of photos depicting doll nipples.[393]

The vulnerability of platforms to public collective action via the media is an important statement on platforms' democratic legitimacy.[394] The media can serve to lend "civility" to individual speech, and render it more capable of effecting change.[395] Though, of course, Facebook, Twitter, and YouTube are not democratic institutions, they arise out of a democratic culture. Thus, users' sense that these platforms respond to collective publicized complaints can impact their trust and use of the company. In a recent survey of Americans who received some of their news from social media, eighty-seven percent used Facebook and trusted the platform more than YouTube and Twitter.[396] These numbers held even following reports that Facebook used politically biased algorithms to post news in its "Trending Topics."[397] While it is impossible to attribute Facebook's high user base and trust entirely to its responsiveness to

---

[389] Amanda Holpuch, *Facebook Adjusts Controversial 'Real Name' Policy in Wake of Criticism*, THE GUARDIAN (Dec. 15, 2015, 1:15 PM), https://www.theguardian.com/us-news/2015/dec/15/facebook-change-controversial-real-name-policy [https://perma.cc/6H9Q-GF7B].

[390] Amy Lee, *Facebook Apologizes for Censoring Gay Kiss Photo*, HUFFINGTON POST (Apr. 19, 2011, 10:36 AM), http://www.huffingtonpost.com/2011/04/19/facebook-gay-kiss_n_850941.html [https://perma.cc/9VNK-HECL].

[391] *Facebook Account Suspended over Nude Courbet Painting as Profile Picture*, THE TELEGRAPH (Apr. 13, 2011, 4:20 PM), http://www.telegraph.co.uk/technology/facebook/8448274/Facebook-account-suspended-over-nude-Courbet-painting-as-profile-picture.html [https://perma.cc/UTS6-YVGA].

[392] Alexei Oreskovic, *Facebook Removes Beheading Video, Updates Violent Images Standards*, HUFFINGTON POST (Oct. 22, 2013, 8:40 PM), http://www.huffingtonpost.com/2013/10/22/facebook-removes-beheading_n_4145970.html [https://perma.cc/Z457-LHUR].

[393] Asher Moses, *Facebook Relents on Doll Nipples Ban*, SYDNEY MORNING HERALD (July 12, 2010), http://www.smh.com.au/technology/technology-news/facebook-relents-on-doll-nipples-ban-20100712-106f6.html [https://perma.cc/2PY8-SD7A].

[394] Jürgen Habermas, *Political Communication in Media Society: Does Democracy Still Enjoy an Epistemic Dimension? The Impact of Normative Theory on Media Research*, 16 COMM. THEORY 411, 419 (2006).

[395] *See* Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 249 (1974) (describing how the press is "enormously powerful and influential in its capacity to manipulate popular opinion and change the course of events"); Robert Post, *Participatory Democracy as a Theory of Free Speech: A Reply*, 97 VA. L. REV. 617, 624 (2011) ("Public opinion could not create democratic legitimacy if it were merely the voice of the loudest or the most violent. . . . Public opinion can therefore serve the cause of democratic legitimacy only if it is at least partially formed in compliance with the civility rules that constitute reason and debate.").

[396] *How People Decide What News to Trust on Digital Platforms and Social Media*, AM. PRESS INST. (Apr. 17, 2016, 10:30 AM), https://www.americanpressinstitute.org/publications/reports/survey-research/news-trust-digital-social-media/ [https://perma.cc/SUZ8-4X9D].

[397] Russell Brandom, *After Trending Topics Scandal, Users Still Mostly Trust Facebook*, THE VERGE (May 18, 2016, 6:00 AM), http://www.theverge.com/2016/5/18/11692882/facebook-public-opinion-poll-trending-topics-bias-news [https://perma.cc/768D-4PP6].

media outcry, the platform's unique history of altering its policies in response to such complaints has likely fostered its user base.

Though ideally democratic, the media can work within this pluralist system to disproportionately favor people with power[398] over the individual users. A series of recent events demonstrate this concept. In September 2016, a well-known Norwegian author, Tom Egeland, posted a famous and historical picture on his Facebook page. The photo of a nine-year-old Vietnamese girl running naked following a napalm attack ("Napalm Girl") was a graphic but important piece of photo journalism from the Vietnam War.[399] It also violated the terms of service for Facebook.[400] The photo was removed, and Egeland's account was suspended.[401] In reporting on the takedown, Espen Egil Hansen, the editor-in-chief and CEO of *Aftenposten*, a Norwegian newspaper, also had the picture removed.[402] Norwegian Prime Minister Erna Solberg also posted the image and had it removed.[403] In response, Hansen published a "letter" to Zuckerberg on *Aftenposten*'s front page. The letter called for Facebook to create a better system to prevent censorship.[404] Hours later, COO Sheryl Sandberg stated that the company had made a mistake and promised the rules would be rewritten to allow the photo.[405] The responsiveness of Facebook would have been more admirable if this had been the first instance of the Napalm Girl photo ever being censored on the site. But instead, it was likely only one of thousands of times the photo had been removed.[406] To the best of my knowledge, however, all prior instances had failed to happen to a famous author, political world

---

398 By this I mean power in every sense: power from money, political clout, media access, access to people that work at platforms, celebrity status, a substantial number of followers or friends, or as a verified user.

399 Kate Klonick, *Facebook Under Pressure*, SLATE (Sept. 12, 2016, 2:48 PM), http://www.slate.com/articles/technology/future_tense/2016/09/facebook_erred_by_taking_down_the_napalm_girl_photo_what_happens_next.html [https://perma.cc/6A4U-UYC5].

400 The photo was likely removed because of the nudity, not because it was child pornography. *See* Kjetil Malkenes Hovland & Deepa Seetharaman, *Facebook Backs Down on Censoring "Napalm Girl" Photo*, WALL ST. J. (Sept. 9, 2016, 3:07 PM), http://on.wsj.com/2bYZtNR [https://perma.cc/SP8M-UQ5D].

401 *Id.*

402 *See* Espen Egil Hansen, *Dear Mark. I Am Writing This to Inform You that I Shall Not Comply with Your Requirement to Remove This Picture.*, AFTENPOSTEN (Sept. 8, 2016, 9:33 PM), https://www.aftenposten.no/meninger/kommentar/i/G892Q/Dear-Mark-I-am-writing-this-to-inform-you-that-I-shall-not-comply-with-your-requirement-to-remove-this-picture [https://perma.cc/49QW-EDUT].

403 Hovland & Seetharaman, *supra* note 400.

404 *See* Hansen, *supra* note 402.

405 Claire Zillman, *Sheryl Sandberg Apologizes for Facebook's "Napalm Girl" Incident*, TIME (Sept. 13, 2016), http://time.com/4489370/sheryl-sandberg-napalm-girl-apology [https://perma.cc/Z7N4-WA2P].

406 Online Chat with Dave Willner, Former Head of Content Policy, Facebook (Sept. 10, 2016).

leader, or the editor-in-chief of a newspaper — and thus, the content had never been reinstated.

Sometimes the speech of powerful people is not just restored upon removal; it is kept up despite breaking the platform policies. In late October, a source at Facebook revealed that Zuckerberg held a Town Hall meeting with employees to discuss why many of then-candidate Donald Trump's more controversial statements had not been removed from the site even though they violated the hate speech policies of the company.[407] "In the weeks ahead, we're going to begin allowing more items that people find newsworthy, significant, or important to the public interest — even if they might otherwise violate our standards," senior members of Facebook's policy team wrote in a public post.[408] Despite that, many employees continued to protest that Facebook was unequally and unfairly applying its terms of service and content-moderation rules.

3. Third-Party Influences. — For a number of years, platforms have worked with outside groups to discuss how best to construct content-moderation policies. One of the first such meetings occurred in 2012, when Stanford Law School invited many of these platforms to be part of a discussion about online hate speech.[409] In April of that year, roughly two dozen attendees — including ask.fm, Facebook, Google, Microsoft, Quizlet, Soundcloud, Twitter, Whisper, Yahoo, and YouTube[410] — met to discuss the "challenge of enforcing . . . community guidelines for free speech" between platforms that have "very different ideas about what's best for the Web."[411] The best practices that came out of these meetings were issued at the conclusion of months of meetings of the Working Group on Cyberhate and were published on the Anti-Defamation League's (ADL) website in a new page called "Best Practices for Responding to Cyberhate" in September 2014.[412] The page "urge[d] members of the Internet Community, including providers, civil society, the legal community and academia, to express their support for this effort and to publicize their own independent efforts to counter cyberhate."[413]

Civil society and third-party groups had and continue to have an impact on the policies and practices of major social media platforms.

---

407 Deepa Seetharaman, *Facebook Employees Pushed to Remove Trump's Posts as Hate Speech*, WALL ST. J. (Oct. 21, 2016, 7:43 PM), http://on.wsj.com/2ePTsoh [https://perma.cc/CH3B-TXF2].

408 *Id.*

409 Rosen, *supra* note 176.

410 Despite the anonymity, the make-up of the group can be estimated from those industry members that signed the best practices at the culmination of the workshops. *See Best Practices for Responding to Cyberhate*, ANTI-DEFAMATION LEAGUE, http://www.adl.org/combating-hate/cyber-safety/best-practices/ [https://perma.cc/KHS4-PZKE].

411 Rosen, *supra* note 176.

412 Press Release, Anti-Defamation League, ADL Releases "Best Practices" for Challenging Cyberhate (Sept. 23, 2014), https://www.adl.org/news/press-releases/adl-releases-best-practices-for-challenging-cyberhate [https://perma.cc/XU3D-MAPZ].

413 *Best Practices for Responding to Cyberhate*, *supra* note 410.

Sit-downs and conversations sponsored by groups like ADL have pushed the creation of industry best practices. Influence also occurs on a smaller scale. "We have a relationship with them where if we flag something for them, they tend to know that it's serious, that they should look sooner rather than later," stated a member of one third-party anti-hate speech group speaking anonymously.[414] But such a relationship isn't exclusive to just organized advocates or established groups. Reporter and feminist Soraya Chemaly recounts directly emailing Sandberg in 2012 regarding graphic Facebook pages about rape and battery of women. "She responded immediately," says Chemaly, "and put us in touch with the head of global policy."[415] Facebook actively encourages this type of engagement with civil society groups, government officials, and reporters. "If there's something that the media or a government minister or another group sees that they've reported and we haven't taken it down, we want to hear about it," said Bickert.[416] "We've been very proactive in engaging with civil society groups all over the world so that we can get a better understanding of the issues affecting them."[417]

In terms of impacting policy, the Working Group on Cyberhate, which was formed in 2012 by the Inter-Parliamentary Coalition for Combating Anti-Semitism and the group of industry leaders and stakeholders at Stanford,[418] continues to exert influence on the platforms. The group regularly meets to try to tailor platform guidelines to strike the correct balance between freedom of expression and user safety.[419] Other groups, like the Electronic Frontier Foundation (EFF), have a slightly less amicable working relationship with these platforms and exist as more like watchdogs than policy collaborators. Launched in 2012, EFF's site, onlinecensorship.org, works to document when user content is blocked or deleted by providing an online tool where users can report such incidents.[420] "Onlinecensorship.org seeks to encourage companies to operate with greater transparency and accountability toward their users as they make decisions that regulate speech," states the site's

---

414 Telephone Interview with T.K. (Jan. 26, 2016) (on file with author).

415 Telephone Interview with Soraya Chemaly, Director, Women's Media Ctr. Speech Project (May 28, 2016); *see also* Christopher Zara, *Facebook Rape Campaign Ignites Twitter: Boycott Threats from #FBrape Get Advertisers' Attention*, INT'L BUS. TIMES (May 24, 2013, 4:26 PM), http://www.ibtimes.com/facebook-rape-campaign-ignites-twitter-boycott-threats-fbrape-get-advertisers-1278999 [https://perma.cc/A5VT-TKCA].

416 Telephone Interview with Monika Bickert & Peter Stern, *supra* note 279.

417 *Id.*

418 *Best Practices for Responding to Cyberhate*, *supra* note 410.

419 ABRAHAM H. FOXMAN & CHRISTOPHER WOLF, VIRAL HATE: CONTAINING ITS SPREAD ON THE INTERNET 120–21 (2013).

420 *Who We Are*, ONLINECENSORHSHIP.ORG, https://onlinecensorship.org/about/who-we-are [https://perma.cc/F2L2-YQH6].

About Page.[421] "By collecting these reports, we're . . . looking . . . to build an understanding of how the removal of content affects users' lives. Often . . . the people that are censored are also those that are least likely to be heard. Our aim is to amplify those voices and help them to advocate for change."[422] The recent, largely opaque, cooperation between content platforms and government to moderate speech related to terrorism is also an issue of concern for EFF, which has urged such groups "not to 'become agents of the government.'"[423] EFF's director of International Freedom of Expression, Jillian York, said, "I think we have to ask if that's the appropriate response in a democracy."[424] "While it's true that companies legally can restrict speech as they see fit, it doesn't mean that it's good for society to have the companies that host most of our everyday speech taking on that kind of power."[425]

4. Change Through Process. — Beyond outside influences, much of the change in moderation policy and guidelines comes simply from the process of moderation. As new situations arise during moderation, platforms will both tweak current policy as well as develop new rules. "People will do everything on the internet," said Jud Hoffman.[426] "Every day you will encounter something new. . . . The difficulty was making sure we were [reacting] fast enough to address the immediate situations that were causing us to consider [changing our approach], but also being thoughtful enough that we weren't flip-flopping on that particular issue every week."[427] Once the team had come to a conclusion about the "trade-offs" for a new policy, the additions would be disseminated in the new guidelines, which would then be distributed as updates to moderators.[428] Many of these judgments continue to be difficult to make, such as, for example, Nicole Wong's story of removal from YouTube of the beating of an Egyptian dissident. The video was restored once its political significance was understood. "You might see an image that at first blush appears disturbing, yet in many cases it is precisely that sort of power image that can raise consciousness and move people to take action and, therefore, we want to consider very, very seriously the possibility of leaving it up," said Peter Stern, head of the Policy Risk Team at Facebook.[429] "We want people to feel safe on Facebook, but that doesn't always mean they're going to feel comfortable, because they may

---

421 *What We Do*, ONLINECENSORSHIP.ORG, https://onlinecensorship.org/about/what-we-do [https://perma.cc/8AJK-AV4R].

422 *Id.*

423 Andrews & Seetharaman, *supra* note 274.

424 Greenberg, *supra* note 383.

425 *Id.*

426 Telephone Interview with Jud Hoffman, *supra* note 148.

427 *Id.*

428 *Id.*

429 Telephone Interview with Monika Bickert & Peter Stern, *supra* note 279.

be exposed to images that are provocative or even disturbing. We want to leave room for that role to be played as well."[430]

In recent years, Facebook's approach to altering its policy has been less passive than simply waiting for new types of content to filter through the system. "We're trying to look beyond individual incidents where we get criticism, to take a broader view of the fabric of our policies, and make sure that we have mitigated risks arising from our policies as much as we can," said Stern.[431] "This means looking at trends . . . at what people within the company are saying . . . be it reviewers or people who are dealing with government officials in other countries. We regularly take this information and process it and consider alterations in the policy."[432]

D. Within Categories of the First Amendment

In light of this new information about how platforms work, how would the First Amendment categorize online content platforms: are they state actors under *Marsh*, broadcasters under *Red Lion* and *Turner*, or more like newspaper editors under *Tornillo*?

Of these, only finding platforms to be state actors would confer a First Amendment obligation — a result that is both unlikely and normatively undesirable. In finding state action, the Court in *Marsh* was particularly concerned with who regulated the municipal powers, public services, and infrastructure of the company town — the streets, sewers, police, and postal service.[433] Subsequent courts have concluded that these facts bear on whether "the private entity has exercised powers that are 'traditionally the exclusive prerogative of the State.'"[434] This Article has detailed how platforms have developed a similar infrastructure to regulate users' speech through detailed rules, active and passive moderation, trained human decisionmaking, reasoning by analogy, and input from internal and external sources. Yet this similarity, while perhaps moving in a direction which might someday evoke *Marsh*, is not yet enough to turn online platforms into state actors under the state action

---

[430] *Id.*

[431] *Id.*

[432] *Id.*

[433] Marsh v. Alabama, 326 U.S. 501, 502–03 (1946).

[434] Blum v. Yaretsky, 457 U.S. 991, 1005 (1982) (quoting Jackson v. Metro. Edison Co., 419 U.S. 345, 353 (1974)). This test is known as the exclusive public function test. If the private entity does not exercise such powers, a court must consider whether "the private party has acted with the help of or in concert with state officials." McKeesport Hosp. v. Accreditation Council for Graduate Med. Educ., 24 F.3d 519, 524 (3d Cir. 1994). The final factor is whether "[t]he State has so far insinuated itself into a position of interdependence with [the acting party] that it must be recognized as a joint participant in the challenged activity." Krynicky v. Univ. of Pittsburgh, 742 F.2d 94, 98 (3d Cir. 1984) (quoting Burton v. Wilmington Parking Auth., 365 U.S. 715, 725 (1961)).

doctrine.[435] In part, this is because while platforms have an incredible governing system to moderate content and perform a vast number of other services which might someday be considered "municipal," they are far from "exclusive" in their control of these rights.[436] As the presence of three major sites for posting this content demonstrates, Facebook, YouTube, and Twitter do not have sole control over speech generally, only speech on their sites.[437]

The Court's recent ruling in *Packingham*, however, could signal a shift that might change this calculus. If the Court is concerned with questions of access in order to exercise constitutionally protected rights, these sites' ability to remove speakers — and the lack of procedure or transparency in doing so — might be of central importance. Still, finding platforms to be state actors seems a long way off and would require a very expansive interpretation of *Marsh*'s current doctrine. Even should the facts necessary to achieve this interpretation come to pass, the normative implications of such a result make it unlikely. Interpreting online platforms as state actors, and thereby obligating them to preserve the First Amendment rights of their users, would not only explicitly conflict with the purposes of § 230, but would also likely create an internet nobody wants. Platforms would no longer be able to remove obscene or violent content. All but the very basest speech would be explicitly allowed and protected — making current problems of online hate speech, bullying, and terrorism, with which many activists and scholars are concerned, unimaginably worse.[438] This alone might be all that is needed to keep platforms from being categorized as state actors.

If these platforms are not state actors, the question of defining them under the First Amendment becomes more complicated. Considering online content providers to be editors like those in *Tornillo*, for instance,

---

[435] *See, e.g.*, Cable Invs., Inc. v. Woolley, 867 F.2d 151, 162 (3d Cir. 1989) (noting that *Marsh* is "construed narrowly").

[436] For a list of the extensive roles that social media and content providers play in users' lives, see Brief for Petitioner at 18–19, Packingham v. North Carolina, 137 S. Ct. 1730 (2017) (No. 15-1194), arguing that access to online speech is protected First Amendment activity because users rely on the sites to exercise religion, contact government officials, receive public notices, assemble, express themselves through music and art, "[a]nd watch cat videos," *id.* at 19. For the assertion that "access to social networking services is indispensible for full participation in the nation's communicative life," see Amicus Curiae Brief of Electronic Frontier Foundation, Public Knowledge, and Center for Democracy & Technology in Support of Petitioner at 8, *Packingham*, 137 S. Ct. 1730 (No. 15-1194) (capitalization omitted).

[437] *Cf.* Cyber Promotions, Inc. v. Am. Online, Inc., 948 F. Supp. 436, 443–44 (E.D. Pa. 1996) (noting, for purposes of state action, that an advertiser banned from AOL could still reach "members of competing commercial online services," *id.* at 443).

[438] *See generally* BAZELON, *supra* note 105; CITRON, *supra* note 102; Citron & Franks, *supra* note 106; Citron & Norton, *supra* note 104; Franks, *supra* note 103. This consequence is of course conditioned on the continued legality of this type of content.

would grant them special First Amendment protection. While platforms' omnipresent role seems to be moving them beyond the world of "editors," *Packingham*'s new labeling of platforms as "forums" makes dismissing this categorization slightly more difficult. In *Tornillo*, the Court held that a newspaper was "more than a passive receptacle or conduit for news, comment, and advertising. The choice of material to go into a newspaper, and the decisions made as to limitations on the size and content of the paper, and treatment of public issues and public officials — whether fair or unfair — constitute the exercise of editorial control and judgment."[439] Thus, in order not to "dampen[] the vigor and limit[] the variety of public debate,"[440] the Court found the newspaper in *Tornillo* to have rights equivalent to a speaker under the First Amendment.[441] At first blush, this analogy seems appealing. As seen above, like the *Miami Herald*, Facebook, YouTube, and Twitter are not "passive . . . conduit[s] for news, comment, and advertising." These platforms have intricate systems for controlling the content on their sites. For the content that stays up — like a newspaper determining what space to allot certain issues — platforms also have intricate algorithms to determine what material a user wants to see and what material should be minimized within a newsfeed, homepage, or stream. But a central piece is missing in the comparison to an editorial desk: platforms do not actively solicit specific types of content, unlike how an editorial desk might solicit reporting or journalistic coverage. Instead, users use the site to post or share content independently. Additionally, platforms play no significant role — yet[442] — in determining whether content is true or false or whether coverage is fair or unfair. As Willner summarized: "This works like a Toyota factory, not a newsroom."[443] Accordingly, while platforms might increasingly be compared to editors as their presence continues to expand in online discourse, they are still far from constituting editors under *Tornillo*.[444]

Perhaps the increasingly apt analogy is — even though the Court in *Reno* explicitly excluded it — to compare platforms to broadcasters, and then perhaps even to public utilities or common carriers.[445] In *Reno*,

---

439 Miami Herald Publ'g Co. v. Tornillo, 418 U.S. 241, 258 (1974).

440 *Id.* at 257 (citing N.Y. Times Co. v. Sullivan, 376 U.S. 254, 279 (1964)).

441 *Id.* at 258.

442 *See, e.g.*, Shannon Liao, *Facebook Now Blocks Ads from Pages that Spread Fake News*, THE VERGE (Aug. 28, 2017, 2:11 PM), https://www.theverge.com/2017/8/28/16215780/facebook-false-viral-hoaxes-trump-malicious-suspicious [https://perma.cc/UZ3D-2BL6].

443 Klonick, *supra* note 399.

444 It is worth noting that Wong and others frequently referred to platforms as possessing their own First Amendment rights to create the type of platform they wanted. This argument stems from *Tornillo*, but it is more ambitious than any rights currently reflected in the doctrine.

445 *See* Kate Klonick, Opinion, *The Terrifying Power of Internet Censors*, N.Y. TIMES (Sept. 13, 2017), http://nyti.ms/2vU9gu9 [https://perma.cc/3V23-2XHV].

the Court explicitly differentiated the internet from broadcast media because the former lacks scarcity, invasiveness, and a history of government regulation.[446] Excepting the lack of historical regulation around the internet, much has changed online since 1998 in terms of internet scarcity and invasiveness. In the years since *Reno*, the hold of certain platforms has arguably created scarcity — if not of speech generally, undoubtedly of certain mediums of speech that these platforms provide. Certainly too, the internet is now more invasive in everyday life than television is — in fact, today, the internet actively threatens to supplant television and broadcasting,[447] and the rise in smartphones and portable electronic technology makes the internet and its platforms ubiquitous. Perhaps most convincingly, in the underlying *Red Lion* decision, the Court argued that "[w]ithout government control, the medium would be of little use because of the cacaphony [sic] of competing voices, none of which could be clearly and predictably heard."[448] The recent scourge of online fake news, scamming, and spam makes this seemingly anachronistic concern newly relevant.

As for public utilities or common carriers regulation, the argument has long been applied at the most basic level of the internet to answer concerns over possible politicization of internet service providers[449] that act as content-neutral conduits for speech. But this argument fails for platforms, because they are inherently *not* neutral — indeed the very definition of "content moderation" belies the idea of content neutrality. Nevertheless, the "essential" nature of these private services to a public right — and the prominence of a few platforms which hold an increasingly powerful market share — evinces concerns similar to those of the people who are arguing for regulation of telephone or broadband services.

A few other analogies that implicate the First Amendment might also apply, but they all fail to match the scope and scale of the speech happening on online platforms. Platforms' use of rule sets to govern speech is reminiscent of "speech codes" used by universities to constrain the speech rights of the student body. But private universities are not truly full-fledged forums — not in the way that California and New Jersey treat shopping malls,[450] and not in the way that platforms have become forums for global public speech.[451] Forums are incidental to the primary

---

446 Reno v. ACLU, 521 U.S. 844, 868 (1997).

447 *See, e.g.*, *Cutting the Cord*, THE ECONOMIST (July 16, 2016), https://www.economist.com/news/business/21702177-television-last-having-its-digital-revolution-moment-cutting-cord [https://perma.cc/N6HC-QE7K].

448 Red Lion Broad. Co. v. FCC, 395 U.S. 367, 376 (1969).

449 GOLDSMITH & WU, *supra* note 100, at 72–74.

450 *See, e.g.*, PruneYard Shopping Ctr. v. Robins, 447 U.S. 74, 78 (1980).

451 Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017).

role of the university, which is to act as an educational institution.[452] The same is true in examining the ability of homeowners' associations or professional organizations to conscribe the speech of members or individuals.[453] The special purposes of universities, professional organizations, or homeowners' associations — to confer knowledge, protect a professional identity, or create a distinct visual community — are distinct from the motives of online speech platforms. Moreover, the global scale and essential nature of private governance of online speech separate it in kind from the strictures governing individuals within these isolated organizations.

The law reasons by analogy, yet none of these analogies to private moderation of the public right of speech seem to precisely meet the descriptive nature of what online platforms are, or the normative results of what we want them to be. The following Part argues for a new kind of understanding: seeing these platforms' regulation of speech as governance.

IV. THE NEW GOVERNORS

Thinking of online platforms from within the categories already established in First Amendment jurisprudence — as company towns, broadcasters, or editors — misses much of what is actually happening in these private spaces. Instead, analysis of online speech is best considered from the perspectives of private governance and self-regulation.[454]

Analyzing online platforms from the perspective of governance is both more descriptively accurate and more normatively useful in addressing the infrastructure of this ever-evolving private space. Platform governance does not fit neatly into any existing governance model, but

---

[452] For excellent discussions of the role of the university in free speech, see generally ROBERT C. POST, DEMOCRACY, EXPERTISE, AND ACADEMIC FREEDOM: A FIRST AMENDMENT JURISPRUDENCE FOR THE MODERN STATE (2012); J. Peter Byrne, *Academic Freedom: A "Special Concern of the First Amendment*," 99 YALE L.J. 251 (1989); and Robert Post, *The Classic First Amendment Tradition Under Stress: Freedom of Speech and the University* (Yale Law Sch., Public Law Research Paper No. 619, 2017), https://ssrn.com/abstract=3044434 [https://perma.cc/B9NH-YFN6].

[453] *See* Claudia E. Haupt, *Professional Speech*, 125 YALE L.J. 1238, 1241–42 (2016).

[454] *See, e.g.*, PASQUALE, *supra* note 125, at 140–68, 187–218 (arguing that terms of service or contracts are inappropriate or ineffective remedies in an essentially "feudal" sphere, *id.* at 144, and that platforms act as "sovereign[s]" over realms of life, *id.* at 163, 189); Freeman, *supra* note 15, at 636–64 (describing the ability of private firms to self-regulate in areas of public interest with and without government influence); Michael P. Vandenbergh, *The Private Life of Public Law*, 105 COLUM. L. REV. 2029, 2037–41 (2005) (discussing how private actors play an increasing role in the traditional government standard-setting, implementation, and enforcement functions through contracts and private agreements). On the role of voluntary self-regulation by private actors, see NEIL GUNNINGHAM ET AL., SMART REGULATION: DESIGNING ENVIRONMENTAL POLICY 167–70 (1998), which analyzes shortcomings of self-regulation, including lack of transparency and independent auditing, concern that performance is not being evaluated, and absence of real penalties for recalcitrants.

it does have features of existing governance models that support its categorization as governance. As Parts II and III demonstrated, platforms have a centralized body, an established set of laws or rules, ex ante and ex post procedures for adjudication of content against rules, and democratic values and culture; policies and rules are modified and updated through external input; platforms are economically subject to normative influence of citizen-users and are also collaborative with external networks like government and third-party groups. Another way to conceptualize the governance of online speech by platforms comes from administrative law, which has long implicated the motivations and systems created by private actors to self-regulate in ways that reflect the norms of a community.[455] Perhaps most significantly, the idea of governance captures the power and scope these private platforms wield through their moderation systems and lends gravitas to their role in democratic culture.[456] Changes in technology and the growth of the internet have resulted in a "revolution in the infrastructure of free expression."[457] The private platforms that created and control that infrastructure are the New Governors in the digital era.

How does this new concept of private platform governors normatively fit in our hopes and fears for the internet? For decades, legal scholars have moved between optimistic and pessimistic views of the future of online speech and long debated how — or whether — to constrain it.[458] But the details of the private infrastructure of online speech were largely opaque. Does this new information and conception allay or augment scholarly concerns over the future of digital speech and democratic culture?

The realities of these platforms both underscore and relieve some of these fears. For the optimists, interviews with the architects of these

---

[455] *See* Freeman, *supra* note 15, at 666; Michael, *supra* note 15, at 175–76; Michael P. Vandenbergh, *Order Without Social Norms: How Personal Norm Activation Can Protect the Environment*, 99 NW. U. L. REV. 1101, 1116–29 (2005).

[456] Balkin, *supra* note 11, at 2296.

[457] *Id.*

[458] Lessig was an early pessimist about the future of the internet, seeing it as a potential means of regulation and control. He specifically worried about the domination of the internet by commercial forces that could be manipulated and controlled by the state. LESSIG, *supra* note 21, at 71. Boyle, Goldsmith, and Wu had similar concerns about the state co-opting private online intermediaries for enforcement. *See* GOLDSMITH & WU, *supra* note 100; Boyle, *supra* note 100, at 202–04. In contrast, Balkin has been largely optimistic about the growth of the internet, the growth of platforms, and the ability of these new speech infrastructures to enhance the "possibility of democratic culture." Balkin, *supra* note 7, at 46. But recently he too has become concerned about the future of online speech and democracy, arguing that private platforms and government can together regulate online speech with less transparency, disruption, and obtrusion than ever before. *See* Balkin, *supra* note 11, at 2342. Scholars like Citron, Norton, and Franks have instead long argued for working with private platforms to change their policies. *See* BAZELON, *supra* note 105, at 279–89; Citron, *supra* note 102, at 121–25; Citron & Norton, *supra* note 104, at 1468–84; Franks, *supra* note 103, at 681–88; *cf.* Citron & Franks, *supra* note 106, at 386–90 (discussing the need for governments to craft criminal statutes prohibiting the publication of revenge porn).

platform content-moderation systems show how the rules and procedures for moderating content are undergirded by American free speech norms and a democratic culture.[459] These ideas are also part of their corporate culture and sense of social responsibility. But perhaps more compellingly, platforms are economically responsive to the expectations and norms of their users. In order to achieve this responsiveness, they have developed an intricate system to both take down content their users don't want to see and keep up as much content as possible. To do this has also meant they have often pushed back against government requests for takedown.[460] Procedurally, platform content-moderation systems have many similarities to a legal system. Finally, platforms have a diverse pluralistic group of forces that informs updates of their content-moderation policies and procedures.

Not only is governance the descriptively correct way to understand platform content moderation, but it is also rhetorically and normatively correct. Historically, speech regulation has followed a dyadic model: a territorial government, with all the power that that invokes, has the boot on the neck of individual speakers or publishers.[461] The New Governors are part of a new model of free expression: a triadic model.[462] In this new model, online speech platforms sit between the state and speakers and publishers. They have the role of empowering both individual speakers and publishers (as well as arguably minimizing the necessity of publishers to speaking and amplification), and their transnational private infrastructure tempers the power of the state to censor. These New Governors have profoundly equalized access to speech publication, centralized decentralized communities, opened vast new resources of communal knowledge, and created infinite ways to spread culture. Digital speech has created a global democratic culture,[463] and the New Governors are the architects of the governance structure that runs it.

The system that these companies have put in place to match the expectations of users and to self-regulate is impressively intricate and responsive. But this system also presents some unquestionable downsides that grow increasingly apparent. These can be seen in two main concerns: (1) worries over loss of equal access to and participation in speech on these platforms; and correspondingly (2) lack of direct platform accountability to their users.

---

459 This is good news for Lessig, Balkin, and Benkler, given their concerns.

460 If this trend continues, it allays much of Balkin's concern over collateral censorship in *Old-School/New-School Speech Regulation*. *See* Balkin, *supra* note 11.

461 Balkin, *supra* note 362 (manuscript at 4, 41).

462 *Id.* (manuscript at 41–44). Balkin refers to this as a "pluralist" model, *id.* (manuscript at 4), and while that term is perhaps more accurate for the world of internet speech as a whole, for my focus here I prefer to use the term "triadic."

463 *Id.* (manuscript at 41–44).

A. Equal Access

There is very little transparency from these private platforms, making it hard to accurately assess the extent to which we should be concerned about speech regulation, censorship, and collateral censorship.[464] But separate from the question of secret government interference or collusion, private platforms are increasingly making their own choices around content moderation that give preferential treatment to some users over others.[465] The threat of special rules for public figures or newsworthy events[466] crystallizes the main value we need protected within this private governance structure in order to maintain a democratic culture: fair opportunity to participate.

In some ways, an ideal solution would be for these platforms to put their intricate systems of self-regulation to work to solve this problem themselves without regulatory interference. But the lack of an appeals system for individual users and the open acknowledgment of different treatment and rule sets for powerful users over others reveal that a fair opportunity to participate is not currently a prioritized part of platform moderation systems. In a limited sense, these problems are nothing new — they are quite similar to the concerns to democracy posed by a mass media captured by a powerful, wealthy elite.[467] Before the internet, these concerns were addressed by imposing government regulation on mass media companies to ensure free speech and a healthy democracy.[468] But unlike mass media, which was always in the hands of an exclusive few, the internet has been a force for free speech and democratic participation since its inception.[469] The internet has also made speech less expensive, more accessible, more generative, and more interactive than it had arguably ever been before. These aspects of online speech have led to the promotion and development of democratic culture, writes Balkin, "a form of social life in which unjust barriers of rank

---

464 These are the concerns expressed by Balkin, Lessig, Tushnet, and Wu. *See* Balkin, *supra* note 11, at 2308–14; LESSIG, *supra* note 21, at 327–29; Tushnet, *supra* note 263, at 1002–15; Wu, *supra* note 61, at 317–18.

465 In September 2017, Twitter announced that it had a different content-moderation rule set for removing President Trump's tweets. Arjun Kharpal, *Why Twitter Won't Take Down Donald Trump's Tweet Which North Korea Called a "Declaration of War,"* CNBC (Sept. 26, 2017, 2:56 AM), https://www.cnbc.com/2017/09/26/donald-trump-north-korea-twitter-tweet.html [https://perma.cc/LXQ6-LXB9]. In December 2015, Facebook similarly disclosed that it had a different set of rules for removing the speech of then-candidate Trump than it had for other users. Doug Bolton, *This Is Why Facebook Isn't Removing Donald Trump's "Hate Speech" from the Site*, INDEPENDENT (Dec. 15, 2015, 6:39 PM), http://www.independent.co.uk/life-style/gadgets-and-tech/news/donald-trump-muslim-hate-speech-facebook-a6774676.html [https://perma.cc/XX4B-CX3V].

466 It is important to note that the uses of "public figure" and "newsworthiness" here differ from their meanings in the sense of communications or privacy torts.

467 Balkin, *supra* note 7, at 30.

468 *Id.* at 31.

469 *See* BENKLER, *supra* note 98; LESSIG, *supra* note 21; Balkin, *supra* note 7, at 3–6.

and privilege are dissolved, and in which ordinary people gain a greater say over the institutions and practices that shape them and their futures. What makes a culture democratic, then, is not democratic *governance*, but democratic *participation*."[470]

Equal access to platforms is thus both an effect of a self-regulated and open internet and the cause of it, making regulation of this issue particularly difficult and paradoxical. Legislating one user rule set for all not only seems logistically problematic, but it would also likely reduce platforms' incentives to moderate *well*. Such legislation, if consitutionally valid, would certainly run into many of the concerns raised by those who fear any regulation that might curb the robust power of § 230 immunity. This is why any proposed regulation — be it entirely new laws or modest changes to § 230[471] — should look carefully at how and why the New Governors *actually* moderate speech. Such, if any, regulation should work with an understanding of the intricate self-regulatory structure already in place in order to be the most effective for users.

B. Accountability

Even without issues of equal access to participation, the central difficulty in simply allowing these systems to self-regulate in a way that takes into account the values and rights of their users is that it leaves users essentially powerless. There is no longer any illusion about the scope and impact of private companies in online platforms and speech.[472] These platforms are beholden to their corporate values, to the foundational norms of American free speech, and to creating a platform where users will want to engage. Only the last of these three motivations for moderating content gives the user any "power," and then only in an indirect and amorphous way.

Moreover, while it initially seems like a positive source of accountability that these systems are indirectly democratically responsive to users' norms, it also creates inherently undemocratic consequences.[473] At first, adaptability appears to be a positive attribute of the system: its ability to rapidly adapt its rules and code to reflect the norms and values of users. But that feature has two bugs: in order to engage with the most users, a platform is (1) disincentivized to allow antinormative content, and (2) incentivized to create perfect filtering to show a user only content

---

[470] Balkin, *supra* note 7, at 35.

[471] *See, e.g.*, Danielle Keats Citron & Benjamin Wittes, *The Internet Will Not Break: Denying Bad Samaritans § 230 Immunity*, 86 FORDHAM L. REV. 401, 414–19 (2017) (proposing limited and narrow revisions to § 230 in order to "not break" the internet).

[472] This was a central concern of Lessig's — that the internet would be captured by large corporations. *See generally* LESSIG, *supra* note 21.

[473] For an excellent discussion of this interplay between corporate power, inequitable markets, and democratic capacity of citizens and users, see generally K. SABEEL RAHMAN, DEMOCRACY AGAINST DOMINATION (2017).

that meets her tastes. These problems are interchangeably known as the so-called echo-chamber effect, which creates an antidemocratic space in which people are shown things with which they already associate and agree, leading to nondeliberative polarization. "It has never been our ideal — constitutionally at least — for democracy to be a perfect reflection of the present temperature of the people."[474] Whether through algorithmic filtering or new content rules, as platforms regress to the normative mean, users will not only be exposed to less diverse content, but they will also be less able to post antinormative content as external and internal content-moderation policies standardize across platforms.

Since the 2016 American presidential election, the lack of accountability of these sites to their users and to the government in policing fake news,[475] commercial speech,[476] or political speech[477] has come to the fore of public consciousness. In statements directly following the election of Trump as President, Zuckerberg emphatically denied the role of fake news in the result.[478] But due to many of the factors discussed here — media pressure, corporate responsibility, and user expectations — Facebook was forced to start tackling the issue.[479] Yet the power of these new threats to "spread[] so quickly and persuade[] so effectively" might make these indirect systems of accountability unexpectedly slow

---

474 LESSIG, *supra* note 21, at 331.

475 Fake news comes in many forms and has notoriously been difficult to define. *See* Claire Wardle, *Fake News. It's Complicated.*, MEDIUM: FIRST DRAFT (Feb. 16, 2017), http://medium.com/1st-draft/fake-news-its-complicated-dof773766c79 [http://perma.cc/EJ9Y-EP6V].

476 Most notably, in late 2017 it was revealed that hundreds of thousands of dollars in ads placed on Facebook during the election had actually come from Russia-linked groups. *See* Mike Isaac & Scott Shane, *Facebook's Russia-Linked Ads Came in Many Disguises*, N.Y. TIMES (Oct. 2, 2017), http://nyti.ms/2g4eVIj [https://perma.cc/SES8-X72P]; Carol D. Leonnig et al., *Russian Firm Tied to Pro-Kremlin Propaganda Advertised on Facebook During Election*, WASH. POST (Sept. 6, 2017), http://wapo.st/2C4pdoH [https://perma.cc/BFK8-HSPW].

477 Following the Russia-linked ads, many platforms have been moving to police more heavily all ad content relating to important issues of political speech. *See, e.g.*, Erik Schelzig, *Twitter Shuts Down Blackburn Campaign Announcement Video*, AP NEWS (Oct. 9, 2017), https://apnews.com/0d8828bd7d204b40af61172628d0a7f6 [https://perma.cc/U97N-37E5] (describing how Twitter blocked an ad by Republican Representative Marsha Blackburn, who was running for the seat being opened by the retirement of Tennessee Senator Bob Corker, in which she boasted that she "stopped the sale of baby body parts," and reporting a Twitter representative's statement that the ad was "deemed an inflammatory statement that is likely to evoke a strong negative reaction").

478 *See, e.g.*, Olivia Solon, *Facebook's Fake News: Mark Zuckerberg Rejects "Crazy Idea" that It Swayed Voters*, THE GUARDIAN (Nov. 10, 2016, 10:01 PM), https://www.theguardian.com/technology/2016/nov/10/facebook-fake-news-us-election-mark-zuckerberg-donald-trump [https://perma.cc/PKD5-BHRW].

479 JEN WEEDON ET AL., INFORMATION OPERATIONS AND FACEBOOK (2017), https://fbnewsroomus.files.wordpress.com/2017/04/facebook-and-information-operations-v1.pdf [https://perma. cc/H9DY-MLHH]; *see also* Carla Herreria, *Mark Zuckerberg: "I Regret" Rejecting Idea that Facebook Fake News Altered Election*, HUFFINGTON POST (Sept. 27, 2017, 8:53 PM), https://www.huffingtonpost.com/entry/mark-zuckerberg-regrets-fake-news-facebook_us_59cc2039e4b05063feoeed9d [https://perma.cc/EY7W-PSNA].

for dealing with such emerging threats and issues.[480] It also makes clear that some insertion of traditional government agency functions — such as regulation of commercial speech — when matched with an accurate understanding of how these platforms currently moderate content, could provide a potential answer to such issues of accountability.[481]

The lack of accountability is also troubling in that it lays bare our dependence on these private platforms to exercise our public rights. Besides exit or leveraging of government, media, or third-party lobbying groups, users are simply dependent on the whims of these corporations. While platforms are arguably also susceptible to the whims of their users, this is entirely indirect — through advertising views, not through any kind of direct market empowerment. One regulatory possibility might be a type of shareholder model — but this fails not only because Zuckerberg owns controlling shares of Facebook, but also because shareholder values of maximizing company profits are perhaps not well matched with user concerns over equal access and democratic accountability. One potential nonregulatory solution to this problem would be for these corporations to register as public benefit corporations, which would allow public benefit to be a charter purpose in addition to the traditional maximizing profit goal.[482]

Another avenue would be for platforms to voluntarily take up a commitment to a notion of "technological due process."[483] In this groundbreaking model for best practices in agency use of technology, Citron advocates for a model that understands the trade-offs of "automation and human discretion," protects individuals' rights to notice and hearings, and gives transparency to rulemaking and adjudication.[484] Of course, these private platforms have little motivation to surrender power as in a public benefit corporation, or to adopt the rules and transparency ideas of Citron's technological due process requirements — but they

---

480 Nabiha Syed, *Real Talk About Fake News: Towards a Better Theory for Platform Governance*, 127 YALE L.J.F. 337, 337 (2017).

481 So far private nongovernmental groups have focused on this. For example, ProPublica has launched a browser attachment to help monitor political ads on online platforms. *See* Julia Angwin & Jeff Larson, *Help Us Monitor Political Ads Online*, PROPUBLICA (Sept. 7, 2017, 10:00 AM), https://www.propublica.org/article/help-us-monitor-political-ads-online [https://perma.cc/A35R-WHHR]. For an excellent and complete discussion of how potential regulation or change should take into account the realities of platforms and moderation, see Syed, *supra* note 480.

482 Kickstarter did this in 2015 in order to make its terms, service, and site more transparent, easier to understand, and easier to access. *See* Yancey Strickler et al., *Kickstarter Is Now a Benefit Corporation*, KICKSTARTER: BLOG (Sept. 21, 2015), https://www.kickstarter.com/blog/kickstarter-is-now-a-benefit-corporation [https://perma.cc/TJ8V-SQT9]. *See generally* David A. Hoffman, *Relational Contracts of Adhesion*, U. CHI. L. REV. (forthcoming 2018), https://ssrn.com/abstract=3008687 [https://perma.cc/NVG9-SHMH] (describing Kickstarter's reincorporation as a public benefit corporation).

483 Danielle Keats Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249, 1301 (2008); *see also id.* at 1301–13.

484 *Id.* at 1301.

might if they fear the alternative would result in more restrictive regulation.[485] Should these platforms come under agency regulation, however, the concerns detailed by Citron's notion of technological due process combined with an accurate understanding of how such companies self-regulate will be essential to crafting responsive and accurate oversight.

CONCLUSION

As the Facebook Live video of Philando Castile's death demonstrates, content published on platforms implicates social policy, law, culture, and the world.[486] Yet, despite the essential nature of these platforms to modern free speech and democratic culture, very little is known about *how* or *why* the platforms curate user content. This Article set out to answer these questions. It began with an overview of the legal framework behind private platforms' broad immunity to moderate content. This framework comes from § 230, the purposes of which were both to encourage platforms to be Good Samaritans by taking an active role in removing offensive content and to protect users' rights by avoiding free speech problems of collateral censorship. With this background, this Article explored why platforms moderate despite the broad immunity of § 230. Through interviews with former platform architects and archived materials, this Article argued that platforms moderate content partly because of American free speech norms and corporate responsibility, but most importantly, because of the economic necessity of creating an environment that reflects the expectations of their users.

Beyond § 230, courts have struggled with how to conceptualize online platforms within First Amendment doctrine: as company towns, as broadcasters, or as editors. This Article has argued that the answer to how best to conceptualize platforms lies outside current categories in First Amendment doctrine. Through internal documents, archived materials, interviews with platform executives, and conversations with content moderators, this Article showed that platforms have developed a system of governance, with a detailed list of rules, trained human decisionmaking to apply those rules, and reliance on a system of external influence to update and amend those rules. Platforms are the New Governors of online speech. These New Governors are private self-regulating entities that are economically and normatively motivated to

---

485 The window for using governmental threat to produce a voluntary result might be closing as the scope and power of these companies make them increasingly difficult to regulate. See, for example, Google's lengthy and robust attempts to push back at the European Court of Justice judgment mandating the "Right to Be Forgotten." *The Right to Be Forgotten (Google v. Spain)*, ELECTRONIC PRIVACY INFO. CTR., https://epic.org/privacy/right-to-be-forgotten/ [https://perma.cc/G3XT-AWR4].

486 While Castile's live-streamed death crystallized the conversation around police brutality and racism in America, it is necessary to note that the officer who shot him was ultimately acquitted. *See* Mitch Smith, *Minnesota Officer Acquitted in Killing of Philando Castile*, N.Y. TIMES (June 16, 2017), http://nyti.ms/2C1MkjF [https://perma.cc/8ETE-LLZE].

reflect the democratic culture and free speech expectations of their users. But these incentives might no longer be enough.

The impact of the video of Philando Castile, the public outcry over Napalm Girl, the alarm expressed at the Zuckerberg Town Hall meeting, and the separate Twitter Rules for President Trump all reflect a central concern: a need for equal access to participation and more direct platform accountability to users. These New Governors play an essential new role in freedom of expression. The platforms are the products of a self-regulated and open internet, but they are only as democratic as the democratic culture and democratic particpation reflected in them. Any proposed regulation — be it entirely new laws or modest changes to § 230 — should look carefully at how and why the New Governors *actually* moderate speech. Such, if any, regulation should work with an understanding of the intricate self-regulatory structure already in place in order to be the most effective for users and preserve the democratizing power of online platforms.