Regulating Online Content Moderation

KYLE LANGVARDT*

The Supreme Court held in 2017 that "the vast democratic forums of the Internet in general, and social media in particular," are "the most important places . . . for the exchange of views." Yet within these forums, speakers are subject to the closest and swiftest regime of censorship the world has ever known. This censorship comes not from the government, but from a small number of private corporations—Facebook, Twitter, Google—and a vast corps of human and algorithmic content moderators. The content moderators' work is indispensable; without it, social media users would drown in spam and disturbing imagery. At the same time, content moderation practices correspond only loosely to First Amendment values. Leaked internal training manuals from Facebook reveal content moderation practices that are rushed, ad hoc, and at times incoherent.

The time has come to consider legislation that would guarantee meaningful speech rights in online spaces. This Article evaluates a range of possible approaches to the problem. These include (1) an administrative monitoring-and-compliance regime to ensure that content moderation policies hew closely to First Amendment principles; (2) a "personal accountability" regime handing control of content moderation over to users; and (3) a relatively simple requirement that companies disclose their moderation policies. Each carries serious pitfalls, but none is as dangerous as option (4): continuing to entrust online speech rights to the private sector.

TABLE OF CONTENTS

* Associate Professor, University of Detroit Mercy School of Law. The author thanks Yafeez Fatabhoy for his valuable research assistance. The author further thanks William Araiza, Erin Archerd, Eric Berger, Gus Hurwitz, Kate Klonick, Chris Lund, Helen Norton, Harvey Perlman, Alex Tsesis, Spencer Weber Walter, and Maggie Wittlin.

INTRODUCTION

In 2017, in *Packingham v. North Carolina*, the Supreme Court struck down a state statute that required registered sex offenders to stay off of social networking services, including Facebook, that might bring them into contact with minors.[1] Justice Kennedy's opinion for the Court explicitly placed the Internet, and social media in particular, on a tier of constitutional importance beyond the "street[s]" and "park[s]" where First Amendment values once had their fullest expression.[2] "While in the past there may have been difficulty in identifying the most important places (in a spatial sense) for the exchange of views," Justice Kennedy wrote, "today the answer is clear. It is cyberspace—the 'vast democratic forums of the Internet' in general, and social media in particular."[3]

1. 137 S. Ct. 1730, 1737–38 (2017).

2. *Id.* at 1735; *see also* Hague v. Comm. for Indus. Org., 307 U.S. 496, 515 (1939) ("Wherever the title of streets and parks may rest, they have immemorially been held in trust for the use of the public and, time out of mind, have been used for purposes of assembly, communicating thoughts between citizens, and discussing public questions. Such use of the streets and public places has, from ancient times, been a part of the privileges, immunities, rights, and liberties of citizens.").

3. *Packingham*, 137 S. Ct. at 1735 (quoting Reno v. ACLU, 521 U.S. 844, 868 (1997)).

Yet today, users of social media are subject to a regime of private censorship[4] that was only recently unimaginable. On Facebook, for instance, users who leave a "cruel or insensitive" comment may face a "cruelty checkpoint" in which a moderator asks them to consider removing it; if they persist, their accounts may be closed.[5] Users may face similar consequences for offending Facebook's often inconsistent policies on hate speech or sexual content.[6]

Some of the overreach is relatively inconsequential, as in the case of the man who was suspended from Facebook for posting a picture of a cat in a tiny business suit,[7] or of the strange decision to blacklist photos of the Little Mermaid statue in Copenhagen, Denmark.[8] But the platform's erratic and opaque decision making can have more serious consequences: at the time of this writing, for instance, Facebook was busily suspending the accounts of Rohingya Muslim groups who

4. I use this term in a descriptive rather than a normative sense. By "censorship," I refer to actions that a public or private governing entity takes on a selective basis to delete expressive content or to prevent speakers from engaging in further expression. "Censorship" therefore encompasses certain laudable policies (such as removing nonconsensual pornography) as well as policies that may be viewed as politically oppressive.

5. Nick Hopkins, *Facebook Moderators: A Quick Guide to Their Job and Its Challenges*, GUARDIAN (May 21, 2017, 1:00 PM), https://www.theguardian.com/news/2017/may/21/facebook-moderators-quick-guide-job-challenges [https://perma.cc/3TN6-MEFL] ("For comments that seem cruel or insensitive, moderators can recommend a 'cruelty checkpoint'; this involves a message being sent to the person who posted it asking them to consider taking it down. If the user continues to post hurtful material, the account can be temporarily closed.").

6. On hate speech, for instance, Facebook's public guidelines vaguely condemn speech that "directly attacks people based on their: race, ethnicity, national origin, religious affiliation, sexual orientation, sex, gender, or gender identity, or serious disabilities or diseases." *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards#hate-speech [https://perma.cc/RMS6-PNC4] (last visited Aug. 2, 2017). Under this policy, Facebook removed a post by Black Lives Matter activist Didi Delgado reading, "All white people are racist. Start from this reference point, or you've already failed." *See* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men From Hate Speech but Not Black Children*, PROPUBLICA (Jun. 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/Z48T-H3K3]. Yet other, far stronger posts have been left intact—including Louisiana Congressman Clay Higgins' reprehensible post on the London terror attacks of 2016:

> The free world . . . all of Christendom . . . is at war with Islamic horror. Not one penny of American treasure should be granted to any nation who harbors these heathen animals. Not a single radicalized Islamic suspect should be granted any measure of quarter. Their intended entry to the American homeland should be summarily denied. Every conceivable measure should be engaged to hunt them down. Hunt them, identify them, and kill them. Kill them all. For the sake of all that is good and righteous. Kill them all. -Captain Clay Higgins.

Captain Clay Higgins (@captclayhiggins), FACEBOOK (June 4, 2017), https://www.facebook.com/captclayhiggins/photos/a.655256107910738.1073741829.581436541959362/997878010315211/?type=3&theater [https://perma.cc/SX7W-2JYB]; *see also* Angwin & Grassegger, *supra* note 6.

7. Michael Moore, *Facebook Sharing this Cat Photo Could Get You BANNED*, EXPRESS (Oct. 5, 2016, 6:40 PM), https://www.express.co.uk/life-style/science-technology/717978/facebook-bans-user-sharing-cat-photo [https://perma.cc/G74Z-SZWT].

8. Mark Molloy, *Facebook Accused of Censoring Photo of Copenhagen's Little Mermaid Statue*, TELEGRAPH (Jan. 4, 2016, 8:53 PM), http://www.telegraph.co.uk/news/worldnews/europe/denmark/12081589/Copenhagen-Little-Mermaid-statue-Facebook-accused-of-censoring-photo.html [https://perma.cc/AJV9-MKBD].

were reporting on the ethnic cleansing of their people in Myanmar.[9] When the girlfriend of Philando Castile went to Facebook Live to simulcast his shooting at the hands of the police, Facebook interrupted the video without explanation.[10]

Users puzzled about the bases of these censorship decisions are forwarded to Facebook's brief and vague "Community Standards" guidelines. But the content moderators' actual rulebook, with the exception of a recent press leak, has for whatever reason been handled as a trade secret.[11] This is not the kind of governance one would normally associate with the phrase "democratic forum." The best that can be said for it is that the platform's content moderators seem generally well-meaning and eager to satisfy popular opinion.[12] But that is no substitute for a guaranteed speech right.

That Facebook is not a governmental actor, of course, relieves all formal constitutional concerns about the company's content restriction policies. But if Justice Kennedy is correct that online platforms have displaced streets and parks as "the most important places . . . for the exchange of views," then it remains important to engage with a deeper set of concerns: What counts as speech? Why is speech special? Is all speech valued equally? How much speech-caused harm does a free speech principle require a society to tolerate? Which questions of expressive liberty should be decided through ordinary politics, and which should be decided under pre-political commitments? How much discretion should be allowed to the censor?

These are high, enduring questions for public institutions, not small matters to be entrusted to the in-house counsel of a few giant, young corporations.[13] Yet

9. Julia Carrie Wong, Michael Safi & Shaikh Azizur Rahman, *Facebook Bans Rohingya Group's Posts as Minority Faces 'Ethnic Cleansing,'* GUARDIAN (Sept. 20, 2017 3:02 AM), https://www.theguardian.com/technology/2017/sep/20/facebook-rohingya-muslims-myanmar [https://perma.cc/DZ7B-4N5L].

10. *See* Timothy Karr, *How Censoring Facebook Affects the Fight for Black Lives*, ROOT (Aug. 29, 2016, 2:03 PM), https://www.theroot.com/how-censoring-facebook-affects-the-fight-for-black-live-1790856542 [https://perma.cc/EL83-MXVH] (discussing the temporary interruption of the Philando Castile police shooting video on Facebook Live, which the company attributed to a "glitch," and explaining that the company nevertheless reserves the right to interrupt disturbing Facebook Live videos that glorify violence or risk imminent harm).

11. Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism and Violence*, GUARDIAN (May 21, 2017, 1:00 PM), https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence [https://perma.cc/L9K7-W4FS].

12. *See* Olivia Solon, *To Censor or Sanction Extreme Content? Either Way, Facebook Can't Win*, GUARDIAN (May 23, 2017, 1:00 AM), https://www.theguardian.com/news/2017/may/22/facebook-moderator-guidelines-extreme-content-analysis [https://perma.cc/SMC3-W4RV] ("So many of these policies are at odds with each other . . . . The company's commitment to these things appears to wax and wane depending on public sentiment" (quoting UCLA Professor Sarah T. Roberts)).

13. *See* JEFFREY ROSEN, BROOKINGS, THE DECIDERS: FACEBOOK, GOOGLE, AND THE FUTURE OF PRIVACY AND FREE SPEECH 10 (2011) https://www.brookings.edu/research/the-deciders-facebook-google-and-the-future-of-privacy-and-free-speech/ [https://perma.cc/28HY-PA85] ("At the moment, the person who arguably has more power than any other to determine who may speak and who may be heard around the globe isn't a king, president or Supreme Court justice. She is Nicole Wong, the deputy general counsel of Google, and her colleagues call her 'The Decider.'"). For a thorough account of the moderators' hierarchy within major social media platforms, see Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. (forthcoming 2018).

those corporations' power over public discourse today is so concentrated and far-reaching that it resembles and arguably surpasses state power within its sphere.[14] "In a lot of ways Facebook is more like a government than a traditional company," Mark Zuckerberg has said. "We have this large community of people, and more than other technology companies we're really setting policies."[15] Ensuring that these platforms do not abuse their state-like powers is a public responsibility worthy of a considered political response.

This Article reveals the treacherous path ahead as the public settles on an architecture of free expression online. I begin by demonstrating that online content moderation—by definition a vast system of prior restraint—is at some level indispensable to Internet communications. The resultant need to accommodate some level of content moderation will inform and significantly complicate any public effort to rein in censorial excesses. With this in mind, I then consider a range of policies that policymakers might someday apply to the censorship activities of online speech platforms.

These include:

(1) across-the-board limits on content moderation practices to ensure that platforms respect a legally-defined set of user speech rights;[16]

(2) a requirement that platforms provide users the ability to toggle content restrictions;[17]

(3) mandatory disclosure of content moderation policies and procedures;[18] and

(4) leaving online platforms the right to moderate content as they please.[19]

It is not an appealing menu, and the options only look worse under examination. Every technique for intervention—except, perhaps, for the mild astringent of disclosure and transparency—is heavy-handed, cumbersome, and nightmarishly complicated. I know of nobody who has called for these kinds of policies.

---

14. *See* Anupam Chander, *Facebookistan*, 90 N.C. L. REV. 1807, 1814–15 (2012) ("[I]t is not the size of Facebook as a corporation alone that makes some use the language of nationhood to describe it. What makes Facebook different from so many other corporations, and more like a government, is how it is involved with so many aspects of our lives, including our business relationships, our friendships, and our families . . . . In some ways, Facebook is more involved with intimate aspects of our lives than governments of liberal states."); Miguel Helft, *Facebook Wrestles with Free Speech and Civility*, N.Y. TIMES (Dec. 12, 2010), http://www.nytimes.com/2010/12/13/technology/13facebook.html?_r=0 [https://nyti.ms/2jfvBRz] ("Facebook has more power in determining who can speak and who can be heard around the globe than any Supreme Court justice, any king or any president . . . ." (quoting ROSEN, *supra* note 13 at 10)).

15. Franklin Foer, *Facebook's War on Free Will: How Technology Is Making Our Minds Redundant*, GUARDIAN (Sept. 19, 2017, 1:00 AM), https://www.theguardian.com/technology/2017/sep/19/facebooks-war-on-free-will [https://perma.cc/7TDL-98RK].

16. *See infra* Part II.

17. *See infra* Part III.

18. *See infra* Part IV.

19. *See infra* Part V.

Yet the full implications of the fourth option—the status quo system in which private companies have free rein to design censorship protocols beyond the rule of law—are almost shockingly dystopian when considered from a distance.

Cloudflare CEO Matthew Prince, who made a unilateral and admittedly "arbitrary" decision to withdraw security support from a neo-Nazi website in 2017, recognized the problem in an internal email: "I woke up in a bad mood and decided someone shouldn't be allowed on the Internet . . . . No one should have that power."[20] He is right. If nothing else, the twentieth century's law of free expression established that only the final censor—at that time, the state—was subject to law. Today, a small number of politically-unaccountable technology oligarchs exercise state-like censorship powers without any similar limitation. We have muddled along under this system for a little over a decade. How long do we expect such a system to last before some of the companies fall into authoritarian hands, or before a sheepish CEO succumbs to governmental pressure in a time of national crisis?

The hard fact is that twenty-first century technology poses a new and unprecedented challenge to our generally laissez-faire system of free speech. Adapting that system to the new technological reality without betraying its values should be the central problem of free speech in the twenty-first century. It will require public institutions to develop new theoretical concepts as grand as the ones that emerged from the early dissents of Holmes and Brandeis.[21] To ignore the big questions here, or to look backward for answers, is to cede the field.

I. THE DILEMMA OF THE MODERATORS

Before exploring policy options, it is necessary to understand that the problem of content moderation raises a new kind of difficulty for the freedom of speech. The broad dilemma is this: the Internet makes it easy for bad actors, ranging from trolls to spammers to malicious hackers, to deter or frustrate speech within online channels.[22] Something must be done to mitigate this problem and fortunately, the technology exists to do so. To use that technology, however, is to adopt a

20. Matthew Braga, *After Cracking Down on Neo-Nazis, Tech Companies Wonder Who Should Police Online Hate*, CBC (Aug. 19, 2017, 5:00 AM), http://www.cbc.ca/news/technology/charlottesville-neo-nazis-white-supremacists-tech-hate-1.4253406 [https://perma.cc/LV6R-C6TC].

21. *See generally* David M. Rabban, *The Emergence of Modern First Amendment Doctrine*, 50 U. CHI. L. REV. 1205, 1305-45 (1983).

22. The actors may not necessarily be "bad." Professor Lidsky observes that:

> Studies reveal that speakers are more prone to be profane or abusive when communication is "computer-mediated." The use of the computer imposes a separation between speaker and audience and thus creates a "disinhibiting" effect. This disinhibiting effect is magnified in instances where the speaker believes himself to be anonymous. The disinhibiting effect is both a virtue and vice of online discourse. On one hand, it leads to a discourse that is "uninhibited, robust, and wide-open." On the other, it leads to more profane and abusive speech. As this type of speech becomes more prevalent, and particularly when it targets private citizens rather than government officials, it may deter many citizens from accessing (or allowing their children to access) social media forums. Indeed, profane and abusive speech ultimately may thwart the use of social media as forums for public discourse.

Lyrissa Lidsky, *Public Forum 2.0*, 91 B.U. L. REV. 1975, 2025 (2011) (footnotes omitted).

pervasive system of prior restraints based on snap judgments. Such a system, whether it is condemned as "censorship" or accepted as "content moderation," sits in tension with an American free speech tradition that was founded on hostility toward *ex ante* administrative licensing schemes.

This dilemma never arose in the twentieth century, when communications were less cheap, fast, anonymous, and platform-dependent than they are today. None of these attributes are new in themselves, but the confluence, and in particular the tension between the first three attributes and the fourth, is unprecedented.

The first three attributes—ease, speed, and anonymity—all drive up the volume and substantive virulence of the communication that takes place at a given time. But all of this communication depends on software platforms that are starkly limited in the amount and character of the communications that they can carry before they become overloaded and their usability deteriorates.

Some of the limits on carrying capacity are technical in nature. Think of a distributed denial-of-service attack, for instance, in which millions of hostile computers access the same IP address at once, overwhelming its capacity to respond to friendly computer users attempting bona fide communications with that address.[23]

Closer to the user interface, the bandwidth limitation may be human rather than technical. Imagine your email without spam filtering, or your Facebook feed if it were populated daily with beheading videos and violent pornography. In each instance, some upper limit on human tolerance plausibly threatens to make a communications platform unusable. The moderator's job is to prevent that from happening—to prevent the inexpense, speed, and anonymity of online communications from crashing disastrously against the limits of platform dependency.

Broadly speaking, limitations on platforms' "bandwidth" have always required "moderators" to police speech. The frequency band is a limited platform for broadcasters; the FCC must lease segments of frequency to prevent a tragedy of the commons.[24] Town meetings are limited platforms; if everyone speaks at once, then no one's speech will matter.[25] Physical spaces such as public parks are also limited in their capacity to carry speech; simultaneous unrelated demonstrations, for instance, could conceivably overwhelm a small park as a venue for effective

23. Mindi McDowell, *Understanding Denial-of-Service Attacks*, U.S. COMPUTER EMERGENCY READINESS TEAM (Feb. 6, 2013), https://www.us-cert.gov/ncas/tips/ST04-015 [https://perma.cc/C3FT-5L3Y].

24. *See* Philip J. Weiser & Dale N. Hatfield, *Policing the Spectrum Commons*, 74 FORDHAM L. REV. 663, 666–74 (2005) (summarizing the traditional case for regulation); *see also* Garrett Hardin, *The Tragedy of the Commons*, 162 SCIENCE 1243, 1244–45 (1968) (summarizing the concept of tragedy of freedom of the commons).

25. *See* Ashutosh Bhagwat, *The Democratic First Amendment*, 110 NW. U. L. REV. 1097, 1112 (2016) (discussing Alexander Meiklejohn's seminal book on free speech and noting, "Mieklejohn [sic] chose as his model for democratic self-governance a New England town meeting. Meiklejohn thus envisioned self-governance, and the activities protected by the First Amendment, as part of an organized, moderated event with strict, and strictly enforced, rules of procedure." (citing ALEXANDER MEIKLEJOHN, POLITICAL FREEDOM: THE CONSTITUTIONAL POWERS OF THE PEOPLE 24–25 (1979))).

protest. Nor are privately-owned platforms a new thing; newspapers and broadcasters have always had limited space for guest contributors.[26]

But the danger to free expression is amplified today because contemporary Internet platforms comprehend and mediate a far larger share of communications than was previously possible. Facebook, after all, is a natural monopoly that channels several currents of electronic communication—from the mainstream national press to one-on-one private chats and voice calls—through a single clearinghouse. Carl Miller, research director at the Center for the Analysis of Social Media, therefore likely understated the matter when he told The Guardian that Facebook's content moderation policy "might be the most important editorial guide sheet the world has ever created."[27] Miller also suggested that this "editorial guide sheet" could be more aptly compared to law: "Private companies are doing what we've only really expected constituted officials of sovereign power to do."[28]

If the danger to free expression is more acute today than in the past, so, too, is the need for moderation. First, today's platforms must confront what one Facebook employee has described to me as a surging "beer bong" of fast, cheap, and often pseudonymous postings.[29] This is mostly a new thing. Unlike today, the sheer *volume* of communications rarely threatened to overload pre-Internet "platforms." Even where a platform's "denominator" of capacity was low—as in a small, one-block public park—the "numerator" of throughput stayed low because of the relative effort involved with in-person expressive activities such as leafleting and public protest. Moreover, the communications themselves moved slowly enough that "slow" institutions, such as courts and administrative boards, could enforce the rules against nonanonymous speakers who overstepped the rules. Online platforms, by contrast, must move aggressively and quickly to suppress an enormous volume of unwanted communications in the form of spam.[30]

---

26. *See* Miami Herald Publ'g. Co. v. Tornillo, 418 U.S. 241, 256–57 (1974) ("[A] newspaper is not subject to the finite technological limitations of time that confront a broadcaster but it is not correct to say that, as an economic reality, a newspaper can proceed to infinite expansion of its column space to accommodate the replies that a government agency determines or a statute commands the readers should have available.").

27. Solon, *supra* note 12.

28. *Id.*

29. My source, who has asked to remain anonymous, was describing the "News Feed," the endless scroll of Facebook posts that a user sees after signing in.These posts are selected algorithmically from a pool of around 1500 for the average user and ranked by a proprietary algorithm that demotes posts users are unlikely to find interesting. Few users see more than a few hundred of these posts. *See* Will Oremus, *Who Controls Your Facebook Feed*, SLATE (Jan. 3, 2016, 8:02 PM), http://www.slate.com/articles/technology/cover_story/2016/01/how_facebook_s_news_feed_algorithm_works.html [https://perma.cc/3VSP-PWG2].

30. *Cf. Spam Email Levels at 12-Year Low*, BBC (July 17, 2015), http://www.bbc.com/news/technology-33564016 [https://perma.cc/DV74-BT3A] (celebrating cybersecurity firm Symantec's report that spam, as a proportion of all email messages, fell below fifty percent for the first time since 2003). Spammers are migrating to social media as email spam filtration improves. *See* Heather Kelly, *83 Million Facebook Accounts Are Fakes and Dupes*, CNN (Aug. 3, 2012, 5:27 AM), http://www.cnn.com/2012/08/02/tech/social-media/facebook-fake-accounts/index.html [https://perma.cc/Z89G-W5DH].

A second type of overload—traumatizing content that deters users from using the platform—is also markedly more threatening today than in the past. Before, the lack of anonymity and the relative difficulty of using pre-Internet speech platforms operated as non-legal restraints on the total volume of shocking content. Occasional public protests with shocking content have always been possible, as in the infamous case of the neo-Nazi march on Skokie[31] or the public trolling campaign of the Westboro Baptist Church,[32] but they have been relatively rare because they involve physical effort, cost, personal confrontation, and a high threat of reputational damage. These deterrents have guaranteed that the total volume of in-person shocking content has rarely been high enough to overwhelm traditional fora.[33]Justice Harlan's advice in *Cohen v. California*, that offended bystanders should "avert[] their eyes,"[34] reflected a judgment that the culture of free speech would be impaired, rather than protected, if shocking speech in public were suppressed.

On social media platforms, the calculus may have reversed itself. The possibility of anonymous speech on the Internet, combined with the ease of "one to many" communications, largely removes the normative and practical constraints that made content-shock rare in the twentieth century. Consider the fate of ChatRoulette, a website launched in 2009 that connected anonymous users in video chat sessions with random strangers, and that quickly devolved into a hub for exhibitionist men.[35] Reddit.com, a platform that moderates content relatively lightly, was forced to admit in March 2015 that "we are seeing our open policies stifling free expression; people avoid participating for fear of their personal and family safety."[36]

Shocking content, whether obscene, violent, or harassing, raises difficult problems of categorization that have challenged courts for decades.[37] Yet the speed of

31. *See* Nat'l Socialist Party of Am. v. Vill. of Skokie, 432 U.S. 43, 43–45 (1977).

32. *See* Snyder v. Phelps, 562 U.S. 443 (2011).

33. *But see* Andrew Katz, *Unrest in Virginia: Clashes Over a Show of White Nationalism in Charlottesville Turn Deadly*, TIME (Aug. 13, 2017), http://time.com/charlottesville-white-nationalist-rally-clashes/ [https://perma.cc/PMY4-W8K9] (chronicling the deadly violence that resulted from the 2017 "Unite the Right" rally in Charlottesville, Virginia).

34. Cohen v. California, 403 U.S. 15, 21 (1971) (advising that those exposed to vulgar language printed on criminal defendant's jacket "could effectively avoid further bombardment of their sensibilities simply by averting their eyes").

35. *See* Alexis C. Madrigal, *Chatroulette's Less Creepy Offspring*, ATLANTIC (Nov. 22, 2010), https://www.theatlantic.com/technology/archive/2010/11/chatroulettes-less-creepy-offspring/66883/ [https://perma.cc/Y54V-STY8].

36. 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*, THEVERGE, https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech [https://perma.cc/7ZV6-M8TH] (last visited Apr. 13, 2016). The company's announcement came after the "CelebGate" posting on Reddit of over 100 female celebrities' private photos, and a subsequent survey in which many Reddit users reported that in light of "hateful or offensive content or community," they would not recommend Reddit to a friend. *Id*.

37. On obscenity, see, for example, Miller v. California, 413 U.S. 15 (1973); A Book Named "John Cleland's Memoirs of a Woman of Pleasure" v. Attorney Gen. of Mass., 383 U.S. 413 (1966); Roth v. United States, 354 U.S. 476 (1957). On violence and harassment, see, for example, Snyder v. Phelps,

communications on online platforms makes enforcement through slower and more constitutionally reliable institutions, such as courts, impossible. Instead, at Facebook, thousands of "Tier 3" content moderators in international call centers—"click workers" in the company's internal argot—each work through roughly one item of flagged content every ten seconds.[38] Increasingly, the work is done by algorithms—artificial neural networks that have been trained to achieve a high rate of "accuracy" in categorizing offensive content. Censors both artificial and human struggle with the nuanced judgments they are required to make.[39]

In many ways, the system resembles the Chinese online censorship leviathan Golden Shield—a system in which roughly 100,000 people police Internet use "around the clock" to remove offending materials as quickly as possible.[40] Messages containing blacklisted language do not post; other messages that violate the rules without using taboo language are removed minutes or hours later.[41] Facebook, YouTube, and other online platforms use these same powers daily. The difference is merely that America's private-sector Golden Shield is tempered by Western cultural expectations and a keen eye on the bottom line. But it is hard to blame Facebook for thinking that it must choose between adopting Golden Shield or becoming ChatRoulette.

It is also hard to blame policymakers for leaving this awful dilemma to Facebook and Twitter. But the problem transcends these companies. It arises not from any one particular feature of today's social and search platforms, but from the basic economy, speed, anonymity, and platform-dependence of Internet communications.[42] The moderators' dilemma is, by all indications, a permanent social problem sewn into the logic of the Internet. Congress should therefore recognize

562 U.S. 443 (2011); Brown v. Entm't Merchs. Ass'n, 564 U.S. 786 (2011); NAACP. v. Claiborne Hardware Co., 458 U.S. 886 (1982); Watts v. United States, 394 U.S. 705 (1969).

38. The "call-centers" are typically located in the Philippines, Ireland, Singapore, India, or Eastern Europe. *See* Klonick, *supra* note 13, at 49. High priority cases, such as imminent threats of violence, and cases about which Tier 3 workers disagree, are escalated to "Tier 2." Profoundly difficult or important cases are escalated to "Tier 1," which consists mostly of attorneys and policymakers. *See id.* at 48–49.

39. *See* Aarti Shahani, *From Hate Speech to Fake News: The Content Crisis Facing Mark Zuckerberg*, NAT'L PUB. RADIO (Nov. 17, 2016, 5:02 AM), http://www.npr.org/sections/alltechconsidered/2016/11/17/495827410/from-hate-speech-to-fake-news-the-content-crisis-facing-mark-zuckerberg [https://perma.cc/XKK9L7KW].

40. *See* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 92–95 (2006); E.H., *How Does China Censor the Internet?*, ECONOMIST (Apr. 22, 2013), http://www.economist.com/blogs/economist-explains/2013/04/economist-explains-how-china-censors-internet [https://perma.cc/R73A-HTXZ].

41. HUMAN RIGHTS WATCH, "RACE TO THE BOTTOM": CORPORATE COMPLICITY IN CHINESE INTERNET CENSORSHIP 12–13 (2006), https://www.hrw.org/reports/2006/china0806/china0806webwcover.pdf [https://perma.cc/HZ7L-UJ9Q].

42. The problem may be escapable, at least to a degree, if changes were made to the Internet's deep protocols. Internet anonymity, in particular, was an early design choice, not a technical necessity. *See* Walter Isaacson, *How to Fix the Internet*, ATLANTIC (Dec. 15, 2016), https://www.theatlantic.com/technology/archive/2016/12/how-to-fix-the-internet/510797/ [https://perma.cc/VUL2-9D2D] ("There is a bug in its original design that at first seemed like a feature but has gradually, and now rapidly, been exploited by hackers and trolls and malevolent actors: Its packets are encoded with the address of their destination but not of their authentic origin."). Such changes, of course, would involve many serious and far-reaching tradeoffs, and they could do more harm than good to speech freedoms globally.The online

that the censorship customs we establish today will have far-reaching consequences for not only the practice but also the cultural concept of free speech. Whether we decide to regulate platform censorship or to leave it to the market, the decision should be considered and deliberate—not punted to the private sector.

Any attempt to protect online speakers from oppressive content moderation must simultaneously accommodate the content moderation that makes the Internet's "vast democratic forums" usable—a delicate and difficult balance. The dilemmic logic of content moderation therefore eliminates the possibility of a "clean" libertarian solution to the problem. Instead, any legal approach to the problem will involve a novel and unsettling tradeoff: the more speech-protective the government's policy, the more hands-on the government's approach will need to be.

In the following sections, I describe four broad regulatory possibilities in descending order of aggressiveness and speech-protectiveness. First, I consider a "mandatory limits" model requiring platforms to observe defined free speech standards.[43] I begin here because the mandatory limits model represents my best estimate of what it would take to impose First Amendment-like restrictions on private platforms. As I demonstrate, this model would require a degree of administrative hassle and governmental intrusion that lacks precedent in the law of free speech. Second, I consider a close variation on mandatory limits: a "personal accountability" model that would require platforms to grant autonomy to users over censorship protocols.[44] Third, I consider a model that would require disclosure of censorship standards to the public.[45] This model is more politically realistic than either the mandatory limits or the personal accountability model. But as I argue, there is little reason for confidence that public or market pressures alone can stand in for a legally guaranteed system of speech rights. Finally, I consider the status quo system—one in which privately owned platforms exercise absolute control over their content moderation practices.[46]

II. MANDATORY LIMITS ON CONTENT MODERATION

I begin with the most aggressive approach possible: one in which the government would oversee private content moderators to ensure that they observe some legally defined set of speech rights. This Article does not confront questions about what kinds of speech are "in" or "out"—whether platforms should be permitted to suppress hate speech, for example. Instead, it focuses more broadly on establishing a framework in which the government, rather than the private platform owner, exercises the final power of review.

censorship issues discussed in this Article—to state what is probably obvious—would never justify these kinds of changes on their own.

43. *See infra* Part II.

44. *See infra* Part III.

45. *See infra* Part IV.

46. *See infra* Part V.

In the next subsections, I touch on the First Amendment objections that platform owners might raise to such a project. Setting these aside, I then demonstrate that a rule limiting content moderation practices would have to come from Congress, rather than from the courts, the administrative state, or state law. Finally, I describe how this system would work.

A. FIRST AMENDMENT OBJECTIONS TO LIMITS ON CONTENT MODERATION

Any official move to limit content moderation on social media platforms will be challenged, both in policy discussions and in formal constitutional litigation, as an abridgment of the platform operators' own "speech" rights as editors or curators. In challenging the new law under the First Amendment, the platforms would today occupy the high ground.

First, there is a long-standing consensus among lower courts that software code is a "language," and therefore constitutes speech—either about some kind of math or about whatever the software does, or perhaps speech about itself—for First Amendment purposes.[47] By this logic, almost all regulation of software products is content-discriminatory and should receive strict scrutiny.[48] But courts rarely go so far in practice, which almost certainly means they recognize that the "language" argument is facile and that following it through to its practical conclusions in a technologically-advanced economy would be insane.[49] Still, the platforms can be expected to press this argument aggressively as an opening bluff.[50]

---

47. *See* Kyle Langvardt, *The Doctrinal Toll of "Information as Speech,"* 47 LOY. U. CHI. L.J. 761, 769–75 (2016) (reviewing First Amendment case law); *see also* Sorrell v. IMS Health, Inc., 564 U.S. 552, 571 (2011) (describing a "rule that information is speech").

48. *See IMS Health*, 564 U.S. at 565–66.

49. *See* Kyle Langvardt, *The Replicator and the First Amendment*, 25 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 59, 68–84 (2014) (demonstrating that treating digital blueprints for 3D-printable objects as "speech" would require strict scrutiny review of products regulation). It would come as quite a surprise if the courts were willing to fill such a tall order, regardless whether they held the ostensible position that "code is speech." *See* Kyle Langvardt, *Remarks on 3D Printing, Free Speech, and Lochner*, 17 MINN. J.L. SCI. & TECH. 779, 797 (2016) ("Talk, after all, is cheap; it is one thing to say that 'code is speech,' and another altogether to follow those words through to what they imply in the 3D printing context. I am skeptical that there is much appetite at all among judges for strict scrutiny review in the field of product regulation."); *cf.* Def. Distributed v. U.S. Dep't of State, 838 F.3d 451, 468–72 (5th Cir. 2016) (Jones, J., dissenting) (stating that she would have extended robust speech protections to digital blueprints for 3D-printable handguns).

50. When the Justice Department demanded that Apple unlock the iPhone's cryptography software, it argued that programming the capability to do so would constitute compelled speech. *See* Matthew Panzarino, Apple Files Motion to Vacate the Court Order to Force It to Unlock iPhone, Citing Constitutional Free Speech Rights, TECHCRUNCH (Feb. 25, 2016), https://techcrunch.com/2016/02/25/apple-files-motion-to-dismiss-the-court-order-to-force-it-to-unlock-iphone-citing-free-speech-rights/ [https://perma.cc/DT8D-8J4C]. For another recent parallel, see the Electronic Frontier Foundation's notice-and-comment filing on a 2014 New York State proposal to regulate the digital currency Bitcoin. "While digital currencies are most commonly thought of as means of payment, at their very essence, digital currency protocols are code. And as courts have long recognized, code is speech protected by the First Amendment." MARCIA HOFFMAN, ELEC. FRONTIER FOUND., INTERNET ARCHIVE, & REDDIT, COMMENTS TO THE NEW YORK STATE DEPARTMENT OF FINANCIAL SERVICES ON BITLICENSE 122014), https://www.eff.org/document/bitlicense-comments-eff-internet-archive-and-reddit [https://perma.cc/RV6C-D9AU]; *see also* Press Release, Elec. Frontier Found., EFF, Internet Archive, and Reddit Oppose New York's BitLicense Proposal (Oct. 21, 2014), https://www.

In a more substantial and down-to-earth argument, platforms would cast their content moderation procedures and algorithms[51] as editorial choices analogous to a newspaper's and argue that any regulation of them should receive strict scrutiny.[52] Somewhat less compellingly, they may also claim that content moderation promotes associational interests by setting "community standards."[53]

The government, meanwhile, would seek to have platforms treated in the same manner as cable service providers, who may be required to carry local and educational television channels in the public interest.[54] But when the Supreme Court upheld these requirements for cable operators, it did so on the understanding that they were only "conduit[s] for the speech of others, transmitting it on a *continuous and unedited* basis to subscribers."[55] Content moderators, on the other hand, do a great deal of "editing," which probably puts them outside the cable-operator precedent.[56]

Certain broader doctrinal trends would also seem to favor the platforms in this debate. First, the Supreme Court today tends to treat all forms of protected communication, ranging from data mining to original political speech, with equal weight in the First Amendment balance.[57] Second, the Court tends to view the First Amendment as essentially deregulatory in nature, and dislikes arguments

eff.org/press/releases/eff-internet-archive-and-reddit-oppose-new-yorks-bitlicense-proposal [https://perma.cc/KZ8Q-VBTV].

51. Even if the algorithms *themselves* are not regarded as speech, it does not necessarily follow that the algorithms' *outputs* should not be protected as speech.The best First Amendment analyses would treat the nonhuman origin of those outputs as a red herring and analyze instead the speech itself using conventional First Amendment tools. *See* James Grimmelmann, *Speech Engines*, 98 MINN. L. REV. 868, 932 (2014) ("Algorithms are a red herring."); Lee Tien, *Publishing Software as a Speech Act*, 15 BERKELEY TECH. L.J. 629, 712 (2000) ("Software poses no special First Amendment problems if we resist the impulse to treat speech as a thing. Most of the problems that seem to plague First Amendment coverage of software become tractable once we focus on software acts instead of software *per se*.").

52. *See* Miami Herald Publ'g. Co. v. Tornillo, 418 U.S. 241 (1974) (striking down Florida statute granting a "right of reply" to political candidates personally attacked in newspaper editorials).

53. *See, e.g.*, Boy Scouts of Am. v. Dale, 530 U.S. 640, 644 (2000) (upholding Boy Scouts' refusal to admit gay scoutmasters as "expressive association"); Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 570 (1995) (comparing coordination of a parade to "the presentation of an edited compilation of speech generated by other persons"); Roberts v. U.S. Jaycees, 468 U.S. 609, 622–29 (1984) (weighing "expressive association" interest in networking organization's membership policies).

54. *See* Turner Broad. Sys., Inc. v. FCC, 520 U.S. 180, 224–25 (1997) (upholding under intermediate scrutiny a requirement that cable television service providers "must carry" local and public television stations).

55. *See generally* Turner Broad. Sys., Inc. v. FCC, 512 U.S. 622, 629 (1994) (emphasis added).

56. *See* Jian Zhang v. Baidu.com, Inc., 10 F. Supp. 3d 433, 439–41 (S.D.N.Y. 2014) (distinguishing search engines from cable operators).

57. *See* Snyder v. Phelps, 562 U.S. 443, 452, 454 (2011) (describing the Westboro Baptist Church picketers—of "God Hates Fags" infamy—as addressing "matters of public import" that "[occupy] the highest rung of the hierarchy of First Amendment values" (quoting Connick v. Myers, 461 U.S. 138, 145 (1983))); Sorrell v. IMS Health, Inc., 564 U.S. 552, 565 (2011) (striking down limits on pharmaceutical data mining and noting that the law targets specific "speakers and their messages for disfavored treatment" which "goes even beyond mere content discrimination, to actual viewpoint discrimination" (quoting R.A.V. v. City of St. Paul, 505 U.S. 377, 391 (1992))); *see generally* Kyle Langvardt, *A Model of First Amendment Decision-Making at a Divided Court*, 84 TENN. L. REV. 833, 838–47 (2017) (discussing the principle that "speech is speech").

that lawmakers might regulate one speaker's communication to enhance that of other speakers.[58] Finally, the contemporary Court does not normally dilute the First Amendment's strength simply because the case arises in a business setting.[59]

Those trends, of course, can change. Certain broad doctrinal contours that appear timeless today in fact developed relatively recently and are surprisingly contingent on the Court's partisan fault line.[60] But it goes without saying that any action the government might take to protect speech rights on social networks will depend for its survival on the future state of First Amendment doctrine.

B. THE NEED FOR CONGRESSIONAL ACTION

In this subsection, I demonstrate that Congress alone—not the courts or agencies or the states—has the authority to put mandatory limits on private content moderation.

1. Obstacles to a Constitutional Common Law Solution

To enforce the First Amendment against online platforms, the courts would have to relax the state action doctrine as applied to speech—or at least speech occurring on privately owned online platforms. Such a transformation in the law is not completely unthinkable, but it is nearly so, and it is hard to imagine it occurring at any point in the foreseeable future.

Counting online platforms as state actors would probably require courts to follow some variation on the "quasi-municipality" doctrine of *Marsh v. Alabama*.[61] In *Marsh*, the Court held that an Alabama state court judge violated the First Amendment when he enjoined a Jehovah's Witness pamphleteer from "trespassing" on the grounds of a privately-owned mining town.[62] The Court identified many similarities between company-owned towns and public municipalities and concluded that robust enforcement of laws against private trespass on company-owned property would interfere with the practical exercise of First Amendment rights.[63]

Amalgamated Food Employees Union Local 590 v. Logan Valley Plaza, Inc.,[64] now overruled, was the *Marsh* doctrine's high-water mark. Writing for the Court, Justice Marshall invalidated an injunction barring protestors from trespassing onto the grounds of a privately-held shopping mall.[65] Because the shopping mall

58. *See* Buckley v. Valeo, 424 U.S. 1, 48–49 (1976) ("[T]he concept that government may restrict the speech of some elements of our society in order to enhance the relative voice of others is wholly foreign to the First Amendment."); Miami Herald Publ'g. Co. v. Tornillo, 418 U.S. 241, 244 (1974) (striking down "right of reply" statute requiring newspapers to give evenhanded op-ed space to opposing political candidates); Langvardt, *supra* note 57, at 852–56.

59. *See* Langvardt, *supra* note 57, at 856–62.

60. *Id.* at 862.

61. 326 U.S. 501 (1946).

62. *Id.*

63. *Id.* at 508–09.

64. 391 U.S. 308 (1968), *abrogated by* Hudgens v. NLRB, 424 U.S. 507 (1976).

65. *Id.* at 325.

was the "functional equivalent" of the company town in *Marsh*—a setting that itself was the functional equivalent of the public square—the state court lacked the power to enjoin speech activities taking place there.[66]

In some ways, a case like *Logan Valley* would appear to offer a blueprint for defining online social platforms as state actors. The expressive stakes in *Logan Valley*, if anything, were far lower than those posed by online content moderation on social media. Justice Marshall's opinion in *Logan Valley* freely acknowledged, after all, that the protesters were "free to canvass the neighborhood . . . [and] to picket on the berms outside the mall."[67] These venues provided at least an accessible alternative channel of communication, if an insufficient one. By contrast, the case is much more serious when a speaker is blocked from Facebook or Twitter, each of which is effectively a whole medium.[68] And in another distinction, online platforms are in every sense created *for the purpose* of being open platforms—a point of central importance in cases involving speech on government property.[69] These considerations suggest that *Logan Valley*'s rationale might apply even more urgently in the social media context.

On the other side, however, the shopping center of *Logan Valley* and the company town of *Marsh* are more literally analogous to the streets and parks of *Hague v. Committee for Industrial Organization*.[70] *Marsh* itself draws on similarities such as the presence of a town sheriff, a post office, and so on.[71] It is not clear how much of this is dispositive and how much is rhetorical. But, for what it is worth, the analogy from public municipalities to company towns is more tangible than the analogy from public municipalities to social media.[72]

The more significant difficulty with applying the state action doctrine to the platforms lies in the fact that internet platforms can "evict" unwanted speakers without involving the courts. In *Marsh* and *Logan Valley*, as well as in other cases that have constitutionalized private relations—cases such as those in the *New*

66. *Id.* at 318, 325.

67. *Id.* at 318 ("[U]nlike the situation in *Marsh*, there is no power on respondents' part to have petitioners totally denied access to the community for which the mall serves as a business district.").

68. "Total medium bans" traditionally receive special judicial disfavor. *See* James Weinstein, *Database Protection and the First Amendment*, 28 U. DAYTON L. REV. 305, 332–34 (2002).

69. *See, e.g.*, United States v. Kokinda, 497 U.S. 720, 728–29 (1990) ("[T]he location and purpose of a publicly owned sidewalk is critical to determining whether such a sidewalk constitutes a public forum."); NAACP v. City of Phila., 834 F.3d 435, 441 (3d Cir. 2016) (applying strict scrutiny to city's written policy preventing noncommercial advertisers from advertising at municipal airport, and noting that "designated public forums . . . are properties that have 'not traditionally been regarded as a public forum [but are] intentionally opened up for that purpose,'" and,"[a]s with traditional public forums, content-based restrictions get strict scrutiny" (quoting Pleasant Grove City v. Summum, 555 U.S. 460, 469–70 (2009))).

70. 307 U.S. 496, 515–16 (1939) (noting that although the "privilege of a citizen of the United States to use the streets and parks for communication of views on national questions may be regulated in the interest of all," such privilege "must not, in the guise of regulation, be abridged or denied").

71. Marsh v. Alabama, 326 U.S. 501, 502–03 (1946).

72. Justice Kennedy's dicta in *Packingham* suggests that six members of today's Court are not overly concerned with the particulars of the "streets and parks" analogy. *See* Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017) (likening the Internet and social media to the "quintessential forum" found in the streets and parks).

York Times v. Sullivan line, for instance, which constitutionalized the field of defamation and other speech torts,[73] or *Shelley v. Kraemer*, which barred racially restrictive covenants on property[74]—the state action takes place the moment a state-operated court intervenes. When YouTube takes down a video, by contrast, there is no state intervention. This is because YouTube, unlike a private real-estate owner, has a lawful self-help remedy—in-house content moderation—that is much quicker, cheaper, and more effective than lawsuits.

Constitutionalizing these activities would require a more radical modification of the state action doctrine than any that has come before—either one that makes the platform the state actor because its actions are "affected with a public interest,"[75] or one that regards state action as omnipresent when the state is capable of adjusting parties' rights relative to each other,[76] or perhaps one that simply does away with the state action doctrine altogether in certain types of cases.

There is no indication today that such a sea change in constitutional thought is on the horizon—or even that it would be advisable. It is unclear, moreover, that courts possess the institutional capacity to police excesses of content moderation on their own. As I explain below, any program to regulate online content moderation would almost certainly rely heavily on an administrative agency.

2. Obstacles to Administrative Rulemaking

Given the well-known difficulty of mobilizing Congress, one might ask whether a future president might turn to the FCC to liberalize social media platforms. It was the FCC under President Obama, after all, that adopted and defended the net neutrality rules that prohibit broadband services from engaging in content

---

73. *See, e.g.*, Hustler Magazine, Inc. v. Falwell, 485 U.S. 46 (1988) (intentional infliction of emotional distress); Time, Inc. v. Hill, 385 U.S. 374 (1967) (public disclosure of private facts); N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964) (defamation).

74. 334 U.S. 1 (1948).

75. The Court adopted this theory of state action in *Munn v. Illinois*, 94 U.S. 113, 126 (1876) ("[W]e find that when private property is 'affected with a public interest, it ceases to be *juris privati* only.'"), before rejecting it in the *Civil Rights Cases*, 109 U.S. 3, 17 (1883) ("[C]ivil rights, such as are guaranteed by the Constitution against State aggression, cannot be impaired by the wrongful acts of individuals, unsupported by State authority in the shape of laws, customs, or judicial or executive proceedings.") and again in *Jackson v. Metropolitan Edison Co.*, 419 U.S. 345, 353–54 (1974). *See* Note, *The Supreme Court 1974 Term*, 89 HARV. L. REV. 47, 139–51 (1975).

76. Professors Peller and Tushnet critique the state-action doctrine's application in First Amendment cases:

> In its evaluation of free speech . . . the judiciary limits itself to a *Lochnerian* concept that people have free speech liberty unless the state has burdened free speech through affirmative governmental acts. The effects of background entitlements on the exercise of free speech rights are immunized from constitutional challenge. Or, to put it another way, application of the state action doctrine to the identification of burdens on free speech assumes that free speech opportunities exist in the social field to such a degree that one can conclude that democratic self-governance exists, as long as the legislature has not "affirmatively" acted to restrict such opportunities—but merely "tolerates" restrictions that arise from the background rules of property and contract.

Gary Peller & Mark Tushnet, *State Action and a New Birth of Freedom*, 92 GEO. L.J. 779, 794 (2004).

discrimination.[77] The authority to regulate broadband providers, however, probably does not extend to services such as search engines and social media platforms.

The regulatory path is untenable because the Communications Act of 1934 prohibits common-carrier regulation against providers of "information services."[78] This is why, in *Verizon v. FCC*, the D.C. Circuit struck down the FCC's net neutrality rules, which required broadband service providers to give equal priority to all data packets, no matter their source.[79] The FCC worked around *Verizon* by reclassifying broadband as a "telecommunications service,"[80] and in *U.S. Telecom Ass'n. v. FCC*,[81] the D.C. Circuit upheld the classification under the *Chevron* test as a permissible interpretation of ambiguous statutory language.[82] But this characterization is less plausible when applied to social media platforms. In its order, the FCC relied heavily on a consumer perception that broadband was a common-carrier telecommunications service akin to the telephone.[83] Defining Facebook or Twitter as telecommunications services on the same basis would seem to overextend the argument.[84]

3. Obstacles to State Level Regulation

Finally, one might look to state law as an alternative to a congressional approach. If a state did make a law limiting online platforms' content moderation practices, however, it would face a serious preemption challenge. Section 230 of the Communications Decency Act of 1996[85] is widely known for sheltering online platforms from vicarious liability for users' speech torts.[86] Somewhat less well-known is section 230(c)(2), which protects platforms from civil liability on account of content moderation practices.[87] Section 230(c)(2) was intended as a

77. *See* Simone A. Friedlander, *Net Neutrality and the FCC's 2015 Open Internet Order*, 31 BERKELEY TECH. L.J. 905, 905–26 (2016) (summarizing the FCC's Obama-era net neutrality orders and the litigation around them).

78. "A telecommunications carrier shall be treated as a common carrier under this chapter only to the extent that it is engaged in providing *telecommunications services*, except that the Commission shall determine whether the provision of fixed and mobile satellite service shall be treated as common carriage." 47 U.S.C. § 153(51) (2010) (emphasis added).

79. 740 F.3d 623, 650, 655–56 (D.C. Cir. 2014).

80. Protecting & Promoting the Open Internet, 30 FCC Rcd. 5601 (2015).

81. 825 F.3d 674 (D.C. Cir. 2016).

82. *Id.* at 701–06; *see* Chevron U.S.A., Inc. v. Nat. Res. Def. Council, Inc., 467 U.S. 837 (1984).

83. *See U.S. Telecom*, 825 F.3d at 697–700.

84. The Communications Act defines "telecommunications service" as "the offering of telecommunications for a fee directly to the public, or to such classes of users as to be effectively available directly to the public, regardless of the facilities used." 47 U.S.C. § 153(53) (2010). "Information service," however, is the "offering of a capability for generating, acquiring, storing, transforming, processing, retrieving, utilizing, or making available information via telecommunications, and includes electronic publishing, but does not include any use of any such capability for the management, control, or operation of a telecommunications system or the management of a telecommunications service." *Id.* § 153(24).

85. 47 U.S.C. § 230 (1996).

86. *Id.* § 230(c)(1) ("No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider.").

87. *Id.* § 230(c)(2) (disclaiming liability for interactive computers users on account of "any action voluntarily taken in good faith to restrict access to or availability of material that the provider or user

"Good Samaritan" provision to prevent platforms from assuming new tort liabilities when they took on the job of content moderation.[88] But the statute's plain language, as well as Congress' patent statutory intention to promote content moderation, would seem to preempt any state policy that would punish overzealous content moderation.

There would also be a dormant commerce clause challenge to state-level regulation. Under the balancing test of *Pike v. Bruce Church, Inc.*, courts weigh the burden the regulation places on interstate commerce against the importance of the in-state regulatory objective.[89] A state-level statute regulating online content moderation would create a heavy burden on interstate commerce. That burden would become worse still if the platforms were forced to take a multistate patchwork approach to content moderation.

The importance of the in-state objective, moreover, would be highly debatable in court. The speech rights vindicated by the state statute I have hypothesized, however important, are not constitutional rights according to any conventional interpretation;[90] if they were, there would be no need for the statute. It is also arguable that a state statute would actually undercut the free speech and associational rights of platforms and their in-state users. A state-level effort to regulate online content moderation would therefore likely fail under *Pike*.[91]

C. STATUTORY DESIGN

If the courts, the FCC, and the states are not well positioned to liberalize social platforms, then only Congress remains. In this section, I address the basic questions of design that the drafters of a statute to limit online content moderation would face. I see three major questions: the scope of the law, the definition of the offense, and the mechanics of enforcement.

1. Scope of Application

As an initial matter, policymakers should try to calibrate the law's scope to avoid inhibiting, rather than protecting, online speech. The pliancy of the relevant terms—"platform," "social media," and "speech"—complicates the issue significantly. Nevertheless, it is fair to say that some platforms are better positioned to

considers to be obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable, whether or not such material is constitutionally protected.").

88. *Id.* § 230(c) (noting "[p]rotection for 'Good Samaritan' blocking and screening of offensive material"); *see id.* § 230(b)(4) ("It is the policy of the United States . . . to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material . . . .").

89. 397 U.S. 137, 142 (1970).

90. *See supra* Section II.A.

91. Courts have struck down several state laws directed against online pornography based on a similar combination of First Amendment and Dormant Commerce Clause concerns. See, e.g., Am. Booksellers Found. v. Dean, 342 F.3d 96, 102–04 (2d Cir. 2003) (striking down Vermont law prohibiting online distribution of sexually-explicit materials to minors, and stating that "it [is] likely that the internet will soon be seen as falling within the class of subjects that are protected from State regulation because they 'imperatively demand[] a single uniform rule" (quoting Cooley v. Bd. of Wardens, 53 U.S. 299, 319 (1851))).

call themselves speakers than are others. Operators of online message boards such as TDPRI.com, a forum dedicated to the Fender Telecaster electric guitar, manage a kind of "platform," yet they are undoubtedly curators, and their admonition to steer clear of "sex, drug, political, religion or hate discussion"[92] comes across less as censorship than as an exercise of associational rights. When Facebook or Twitter refer to themselves as curators, on the other hand, the claim is mostly opportunistic;[93] both companies operate general-purpose platforms for discussion that operate at enormous scales. Their rules barring hate speech, for example, approach the kind of governance activity one would expect from a sovereign.[94]

Ideally, then, any statute placing limits on content moderation will be selective in its application.Two expressive interests matter here. First, government intervention is more desirable to the extent that users lack ample alternative channels to speak outside a given platform. Second, intervention is less desirable if platform administrators have a strong speech or associational interest in their content moderation practices.[95] These two interests, though separate, are strongly correlated: if the first interest favors limits on content moderation, the second interest probably does as well, and vice versa.

Market power is the phenomenon that binds user and administrator speech interests into this inverse correlation. Social networks are networks, and as such, they depend on network effects for success.[96] Those with many users are valuable, and those with few users are worthless except for niche purposes.[97] Platforms pursuing the scale needed to exploit the network effect therefore cannot be choosy about the users with whom they "associate." If anything, the market would seem to reward an all-comers policy.

The larger and more eclectic a social network becomes, the more the user is "locked in" to it, as opposed to alternative channels. These locked-in users'

92. *POSTING RULES THAT ALL MUST FOLLOW. PLEASE READ!*, TDPRI.COM (Mar. 5, 2016), http://www.tdpri.com/threads/posting-rules-that-all-must-follow-please-read.616711/ [https://perma.cc/JV8W-HRG2].

93. *See generally* Frederick Schauer, *First Amendment Opportunism*, *in* ETERNALLY VIGILANT: FREE SPEECH IN THE MODERN ERA 174–97 (Lee C. Bollinger & Geoffrey R. Stone eds., 2002) (coining the term "First Amendment opportunism" to describe cases in which litigants seeking victory by any means necessary press tenuous First Amendment claims).

94. *See supra* note 12.

95. *See* Hurley v. Irish-Am. Gay, Lesbian & Bisexual Grp. of Bos., 515 U.S. 557, 570 (1995) (comparing coordination of a parade to "the presentation of an edited compilation of speech generated by other persons"); Roberts v. U.S. Jaycees, 468 U.S. 609, 622–29 (1984) (weighing "expressive association" interest in networking organization's membership policies).

96. *See* Spencer Weber Waller, *Antitrust and Social Networking*, 90 N.C. L. REV. 1771, 1787 (2012) ("Network effects refer to the well-known phenomenon that systems may quickly increase in value as the number of users grow, and similarly, that the network may have little, or no, value without large scale adoption." More specifically, "[d]irect network effects refer to systems like communications networks whose value directly increases as the number of users increase . . . [Direct] network effect[s] can create significant entry barriers . . . ." (citing CARL SHAPIRO & HAL R. VARIAN, INFORMATION RULES: A STRATEGIC GUIDE TO THE NETWORK ECONOMY 13 (1999); Michael L. Katz & Carl Shapiro, *Systems Competition and Network Effects*, 8 J. ECON. PERSP. 93, 109 (1994))).

97. *See id.*; *see also infra* notes 156–59 (discussing the adoption of Gab, a small Twitter alternative, by hard-right fringe users concerned about censorship on Twitter).

speech interest is high. At the same time, the biggest social platforms' indiscriminate appetite for scale in the form of new users and more activity seriously undercuts any claim that their users somehow speak for them.[98] Such platforms' speech interest is low. As such, the megaplatforms that draw the most market power from network effects and lock-ins would seem to be the best candidates for regulation.Assessments of market power would therefore figure prominently, in one form or another, in any program to limit online content moderation.

The ideal approach, theoretically, would be to define the category of actors subject to the new restrictions explicitly by reference to market power. But applying antitrust concepts to social media platforms raises several conceptual difficulties.[99] Platforms offer heavily differentiated products, which makes it difficult to say whether, and in what respects, one platform is in competition with another.[100] Are Facebook and Google, for instance, "competitors"? What is the relevant good in this market? Are we concerned with their power in a market to acquire user data? To acquire "views"? To sell targeted advertising services?[101] The statute could, of course, define market power loosely, with the responsibility of classification delegated to the judiciary or an administrative body. But this kind of approach has its own problems, including doctrinal uncertainty surrounding the concept of market power[102] and the danger of agency capture.

Policymakers may therefore be drawn to a mechanism that would avoid the classification problem by inducing platforms to sort themselves. One tempting, if ultimately inadvisable, source of leverage is found in Section 230 of the Communications Decency Act, which protects "internet service providers" from

98. *See* Chris Anderson, *The Long Tail*, WIRED (Oct. 1, 2004, 12:00 PM), https://www.wired.com/2004/10/tail/ [https://perma.cc/ M346-HNWU] (explaining that large online platforms maximize profits not by targeting the median user, but by reaching *all users*, no matter how marginal their tastes).

99. *See generally* Waller, *supra* note 96; Christopher S. Yoo, *When Antitrust Met Facebook*, 19 GEO. MASON L. REV. 1147 (2012).

100. Spencer Waller has observed that when competition does exist among online platforms, it is a temporary and winner-take-all contest of Schumpeterian "creative destruction." *See* Waller, *supra* note 96, at 1800–04. In this dynamic, each platform creates its own "lane" and occupies it exclusively until, following a brief struggle for dominance, it is decisively overthrown by a successor platform. Facebook, for instance, is the current holder of a lane that was previously occupied by Myspace, which in turn overthrew Friendster. *Id*. Smaller-scale developments, such as Facebook's relatively recent adoption of Skype-like videochat capabilities, can be seen through the same lens.

101. Google and Facebook together take in over half of online ad revenue worldwide and over sixty percent in the United States. *See* Reuters, *Why Google and Facebook Prove the Digital Ad Market Is a Duopoly*, FORTUNE (July 28, 2017), http://fortune.com/2017/07/28/google-facebook-digital-advertising/ [https://perma.cc/N6VF-XDTS].

102. Louis Kaplow observes that antitrust doctrine has converged on a definition of market power as the ability to set prices above the competitive level, but that ability is only one among many forms that market power can take. Louis Kaplow, *On the Relevance of Market Power*, 130 HARV. L. REV. 1303, 1396 (2017) ("Inquiries into the various determinants of market power and related concepts are often of great importance in analyzing allegedly anticompetitive practices, but the uses to which the results are appropriately put often differ markedly from those conventionally advanced in competition law doctrine and commentary."). A definition of market power that is constrained to cases of price-setting is of little use in analyzing major social platforms' market position.

derivative civil liability for the acts of their users.[103] Section 230 is often praised as the Internet's Magna Carta because it has given entrepreneurs and forum hosts room to create new culture and products without assuming vicarious liability for their users' defamatory statements.[104] But Frank Pasquale, among others, has argued that section 230 allows platforms to "have it both ways."[105] In other words, when they want to avoid vicarious liability, they cast themselves as passive conduits for speech; but when they object to regulation, they claim to be editors of speech entitled to robust protections.[106] Pasquale proposes that policymakers should assign one status or the other to online platforms: "policymakers could refuse to allow intermediaries to have it both ways, forcing them to assume the rights and responsibilities of content-provider or conduit."[107]

Another way to sort regulable from unregulable platforms would be to present the choice to the platforms themselves: as a condition for "opting in" to section 230 protections, a platform must assume an obligation to protect users' speech rights as provided in the new law. But this seemingly elegant, Solomonic compromise has two serious problems.

First, the publisher-distributor dichotomy excludes hybrid cases. The notion of a strict dichotomy, however defensible in the pre-Internet media environment, has today melted into a publisher-distributor spectrum. Even if it is disingenuous for Facebook, the edge case, to claim that it is both conduit and speaker, the claim is much more plausible in the case of TDPRI, the small discussion board,[108] or—for that matter—nytimes.com, which houses both a major newspaper and an integrated user discussion forum.[109] These fora, along with most sites on the open web, have both publisher and distributer attributes. Forcing a dichotomous choice on these publisher-distributors can only inhibit genuine speech interests. If they lose their right to moderate content on a discriminatory basis, then they lose the freedom to define themselves; if they give up their section 230 protections, then they face a high risk of tort liability—granted, one that is mitigated substantially by First Amendment limits on speech-based tort claims.[110]

Second, self-sorting may exacerbate the censorship problem it intends to solve. If the alternative to limits on content moderation is to face publisher liability for defamation and other torts, then well-staffed or technologically-advanced

103. 47 U.S.C. § 230(c) (2012).

104. *See, e.g.*, *CDA 230: The Most Important Law Protecting Internet Speech*, ELECTRONIC FRONTIER FOUNDATION, https://www.eff.org/issues/cda230 [https://perma.cc/FL2D-G9QD] (last visited Apr. 8, 2018).

105. Frank Pasquale, *Platform Neutrality: Enhancing Freedom of Expression in Spheres of Private Power*, 17 THEORETICAL INQ. L. 487, 496–97 (2016).

106. *See id.*

107. *Id.*

108. *See supra* note 92.

109. Almost all online newspapers include integrated comment sections alongside most articles. *See, e.g.*, N.Y. TIMES, https://www.nytimes.com.

110. *See supra* note 73.

platforms may brute-force their way out of the dilemma by simply censoring all potentially tortious speech.

For these reasons, it properly falls to government actors rather than to the platforms themselves to determine which platforms' content moderation practices should be subject to oversight. Whether stated explicitly in the statute or read in by courts, market power would have to be an element of the offense. The complexity, ambiguity, and high costs associated with litigating that issue probably counsel in favor of some kind of bright-line shelter for small platforms—for example, an absolute exemption from liability under the statute for platforms boasting fewer than fifty million registered users.[111]

2. Defining the Offense

After deciding which online entities will be subject to limits on content moderation, the next challenge is to define those limits. What, exactly, does the statute forbid, and what can online platforms do to ensure compliance?

Legislatures sometimes write statutes that incorporate specific tests from Supreme Court opinions on the First Amendment and apply them in settings where the Constitution does not reach.[112] But a statute can invoke constitutional values without referring with such high resolution to specific doctrinal features. Anti-SLAPP (Strategic Litigation Against Public Participation) statutes, for instance, deter strategic lawsuits meant to chill the exercise of First Amendment liberties, and they do so in summary language without any specific invocation of First Amendment doctrine.[113] At its core, therefore, a statute to liberalize online content moderation might simply prohibit online platforms with dominant market positions from "interfering" with the exercise of constitutional freedoms of speech, religion, and assembly—or something similarly broad—and leave it at that.

A more specific approach consisting of hornbook-style articulations of First Amendment doctrine would be inadvisable. First, any overly granular statute would diverge over time from First Amendment doctrine as it develops in the

111. I borrow this number from the Honest Ads Act, a bill now under consideration to impose disclosure requirements on online political advertisers. S.1989, 115th Cong. § 8 (2017).

112. The Religious Freedom Restoration Act, which reinstated the then-recently-overruled doctrines of *Sherbert v. Verner*, 374 U.S. 398 (1963), and *Wisconsin v. Yoder*, 406 U.S. 205 (1972), is the most famous example of a law which incorporated a Supreme Court First Amendment test. 42 U.S.C. 2000bb(b)(1) (2012) (announcing intent "to restore the compelling interest test as set forth in *Sherbert v. Verner* . . . and *Wisconsin v. Yoder* [after the Supreme Court overruled them in *Employment Division v. Smith*, 494 U.S. 872 (1990)] and to guarantee its application in all cases where free exercise of religion is substantially burdened"). Virginia's recent statute applying content neutrality doctrine to its college campuses is another such example. *See* VA. CODE. ANN. § 23.1-401 (2016).

113. California provides one example:

> A cause of action against a person arising from any act of that person in furtherance of the person's right of petition or free speech under the United States Constitution or the California Constitution in connection with a public issue shall be subject to a special motion to strike, unless the court determines that the plaintiff has established that there is a probability that the plaintiff will prevail on the claim.

CAL. CIV. PROC. CODE § 425.16(b)(1) (West 2015).

courts; it is better that the law of content moderation have the opportunity to develop in parallel to First Amendment law to the extent that it is possible and desirable. Second, the Internet intermediary context is sufficiently novel and fluid that those responsible for setting limits on content moderation would be forced immediately to break from certain First Amendment principles. The traditional prohibition against prior restraints, for example, cannot be carried over into a set of rules that is designed to permit any substantial amount of electronic content moderation.

Nor should the aim of such statutes be to impose liability for individual content moderation decisions that violate the rules. A rough calculation reveals that about five percent of all daily postings to Facebook are reviewed by the company's moderation team.[114] The commercial content moderator operates in a faster, higher-volume environment than prosecutors and other government actors, and is far more inclined toward censoring questionable content. It would be difficult or impossible for commercial content moderators, no matter how skilled, to implement First Amendment doctrine with high precision—and that is doubly true for the AI content moderators that will eventually take over[115] the human moderators' arduous and trauma-inducing[116] job.[117] Policymakers must therefore recognize (1) that takedowns will be frequent and necessary, and (2) that some degree

114. About 900,000 comments, status updates, and photos post to Facebook every minute. Arif Anik, *The Top 20 Valuable Facebook Statistics—Updated April 2017*, LINKEDIN (Apr. 16, 2017), https://www.linkedin.com/pulse/top-20-valuable-facebook-statistics-updated-april-2017-arif-anik [https://perma.cc/PG2B-39LS]. Many outlets have reported that Facebook content moderators evaluate one post every ten seconds, or six per minute. *See, e.g.*, Aarti Shahani, *From Hate Speech to Fake News: The Content Crisis Facing Mark Zuckerberg*, NAT'L PUB. RADIO (Nov. 17, 2016, 5:02 AM), http://www.npr.org/sections/alltechconsidered/2016/11/17/495827410/from-hate-speech-to-fake-news-the-content-crisis-facing-mark-zuckerberg [https://perma.cc/XKK9-L7KW]. If Facebook employs 7,500 of these workers, as reported by Forbes, then Facebook's content moderation corps must clear roughly 45,000 posts per minute, or five percent of the total volume (45,000 divided by 900,000). *See* Kathleen Chaykowski, *Facebook is Hiring 3,000 Moderators in Push to Curb Violent Videos*, FORBES (May 3, 2017, 11:41 AM), https://www.forbes.com/sites/kathleenchaykowski/2017/05/03/facebook-is-hiring-3000-moderators-in-push-to-curb-violent-videos/ [https://perma.cc/FQ8A-LNM7].

115. *See* Josh Constine, *Facebook Spares Humans by Fighting Offensive Photos With AI*, TECHCRUNCH (May 31, 2016), https://techcrunch.com/2016/05/31/terminating-abuse/ [https://perma.cc/9AJU-W38H] ("[T]oday we have more offensive photos being reported by AI algorithms than by people. The higher we push that to 100 percent, the fewer offensive photos have actually been seen by a human." (quoting Joaquin Candela, Facebook's Director of Engineering for Applied Machine Learning)).

116. June Williams, *Workers on Porn Detail Sue Microsoft for Injuries*, COURTHOUSE NEWS (Jan. 10, 2017), https://www.courthousenews.com/workers-on-porn-detail-sue-microsoft-for-injuries/ ([https://perma.cc/2GCF-5Z5V] (describing the lawsuit brought by two Microsoft content moderators for developing post-traumatic stress disorder).

117. "Skin filters," for instance, mark images containing large numbers of contiguous flesh-toned pixels as pornography if they do not also recognize them as close-up portraits. *See* Sarah T. Roberts, *Social Media's Silent Filter*, ATLANTIC (Mar. 8, 2017), https://www.theatlantic.com/technology/archive/2017/03/commercial-content-moderation/518796/ [https://perma.cc/8TWX-4M8S]. This is a useful approach for identifying nudity, but it is inadequate to identify obscenity or indecency.Perhaps later generations of AI will possess the cultural and emotional awareness to make the necessary judgments; but at that point, there is the question of whether we want the bulk of the common law on speech issues to percolate up from machine judgments.

of human or algorithmic error is inevitable in light of the volume of content under review and the subtlety of the principles governing the status of violent and sexually explicit content.[118]

Any reasonable statutory framework would therefore try to focus judicial, regulatory, and corporate attention on sound content moderation *policies* rather than fussing over individual cases. As elsewhere in the law of online intermediaries, lawmakers would likely be drawn to some sort of safe-harbor approach.[119] For example, platforms might be shielded from statutory liability if they kept clearly articulated and well-drawn content moderation policies, made good faith efforts to follow those policies, and made reasonable efforts to correct clear errors.[120] The devil, of course, is in the details.

3. Enforcement and Safe Harbors

The safe-harbor approach that I have proposed would require courts or an agency—or most likely some combination of both—to oversee the design and implementation of platforms' content moderation policies. In a primarily judicial model, courts would dismiss civil claims brought under the statute if defendant platforms established that their content moderation policies qualified for the safe harbor. In a primarily administrative model, platforms might maintain safe-harbor status on a renewable, periodic basis, much like a trademark owner that renews its registration every ten years.[121]

Each mode of oversight has well-known benefits and drawbacks. Common law dispute resolution is slow, generalistic, and costly, but stable and authoritative. Agency oversight is relatively quick, specific, and technically expert, but is subject to industry capture, prone to overregulation, and potentially vulnerable to executive sabotage.

Realistically, the work of defining standards for content moderation would involve some elements of both judicial and administrative oversight. The "high" questions involving the relationship between traditional speech freedoms and

---

118. It is still an open question, for instance, whether the First Amendment requires subjective intent on the part of a defendant charged with making a true threat. Even where the standards are fairly well defined—for instance, in the area of child pornography—they involve nuance. It was a clear mistake when a Facebook moderator took down the iconic "Napalm Girl" photo of a young Vietnamese girl running naked and panicked from a napalm attack on her village. *See* Sam Levin, Julia Carrie Wong & Luke Harding, *Facebook Backs Down from 'Napalm Girl' Censorship and Reinstates Photo*, GUARDIAN (Sept. 9, 2016, 1:44 PM), https://www.theguardian.com/technology/2016/sep/09/facebook-reinstates-napalm-girl-photo [https://perma.cc/6VXB-HC9P]. But saying *why* it was a mistake is more than a ten-second job.

119. The Digital Millennium Copyright Act, for example, provides a safe harbor against vicarious copyright infringement liability to platforms that "upon notification of claimed infringement . . . respond [] expeditiously to remove, or disable access to, the material that is claimed to be infringing or to be the subject of infringing activity." 17 U.S.C. § 512(c)(1)(C) (2012).

120. I borrow loosely from *Burlington Industries, Inc. v. Ellerth*, 524 U.S. 742, 745–46 (1998), and *Faragher v. City of Boca Raton*, 524 U.S. 775, 790–93 (1998), which established an affirmative defense to vicarious employer liability in cases involving Title VII harassment by supervisors.

121. *See* 15 U.S.C. § 1058(a); 37 C.F.R. § 2.181(a)(1) (providing that mark holders who satisfy defined criteria may insulate the validity of their mark from challenge at ten-year intervals).

new statutory speech freedoms seem ideally suited for judges. The day-to-day "low" questions dealing with the implementation of those standards through algorithms and corporate policy would be better left to an agency-driven monitoring and compliance regime.

As a matter of course, all agency oversight would be subject to some level of judicial review. Lawmakers could tweak the apportionment of responsibilities between agencies and the judiciary by customizing standards of review. And to the extent that Congress is concerned about the agency's vulnerability to regulatory capture or executive sabotage, Congress could clarify in the statute that the agency's enforcement obligation is non-discretionary and that refusals to enforce are judicially reviewable.[122]

Online content moderation policies cannot realistically track judicial tests with precision. Facebook, for instance, would be unwise to use the tests from *Miller v. California*[123] or *Brandenburg v. Ohio*[124] as guidelines for the removal of sexually explicit or violence-inciting content. Those tests were formulated for a different decisionmaking environment, and they are too philosophical for the rapid-fire, low-context nature of the content moderator's job. Instead, platforms are likely to draw up blunt, quick heuristics that rely more on checking off red flags than on weighing the equities. It would be inappropriate and futile to expect content moderation standards to be written differently.[125] Any agency review of content moderation guidelines should therefore concern itself less with perfect alignment with constitutional doctrine than with keeping the error rate at some tolerable level.

This point becomes especially important in the case of algorithmic content moderation. Most algorithms used in content moderation are "black boxes"—their results can be evaluated, but their internal processes are incomprehensible to humans.[126] The algorithms are not "instructed" or "programmed" to process

---

122. *See* Heckler v. Chaney, 470 U.S. 821, 831 (1985) (holding that agency refusals to take enforcement action are presumptively unsuitable for judicial review).

123. The *Miller* test is as follows:

> The basic guidelines for the trier of fact must be: (a) whether "the average person, applying contemporary community standards" would find that the work, taken as a whole, appeals to the prurient interest (b) whether the work depicts or describes, in a patently offensive way, sexual conduct specifically defined by the applicable state law; and (c) whether the work, taken as a whole, lacks serious literary, artistic, political, or scientific value.

413 U.S. 15, 24 (1973).

124. The *Brandenburg* test provides that:

> [T]he constitutional guarantees of free speech and free press do not permit a State to forbid or proscribe advocacy of the use of force or of law violation except where such advocacy is directed to inciting or producing imminent lawless action and is likely to incite or produce such action.

395 U.S. 444, 447 (1969).

125. Kate Klonick describes an evolution in Facebook's content moderation from standards to 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) increased reliance on teams of human moderators with diverse backgrounds." Klonick, *supra* note 13, at 43.

126. *See generally* Maayan Perel & Niva Elkin-Koren, *Black Box Tinkering: Beyond Disclosure in Algorithmic Enforcement*, 69 FLA. L. REV. 181, 188–90 (2017) (discussing the non-transparent nature of algorithms).

censorship decisions according to a set of rules; instead, they are "trained" to classify offending content just as a police dog is trained to sniff for drugs.[127] The algorithm begins in an arbitrary configuration and is then tasked with classifying a large data set. As the algorithm converts individual "inputs"—the content—into "outputs"—predictions as to whether the content is prohibited—it is "rewarded" for accuracy and "punished" for inaccuracy. When punishments occur, the system tweaks some part of the algorithm and tries again. Through trial and error, this method can produce algorithms with high rates of accuracy in applications such as facial recognition. But it is impossible for humans to understand *how* or *why* the algorithms make their judgments; looking at the underlying machine code is less like looking at a human's thought process than it is looking at a map of an animal's neurons. Those neurons' interactions cannot easily be reduced to legalistic "elements" or "factors."

Agencies evaluating these algorithms, then, would not concern themselves with the bases for the algorithms' judgments. Those judgments lie in a black box and cannot be adjusted. Instead, the agencies would likely operate in a "quality control" mode, feeding pseudo-constitutional "problem sets" into the algorithms and evaluating the quality of the results. The agency might even maintain stock algorithms for common problems, such as sorting out threatening language, and make them available to firms whose in-house algorithms were found to have unacceptable failure rates.

D. THE CONSEQUENCES OF MANDATORY LIMITS

To summarize, a legal limit on content moderation standards may help to ensure meaningful speech rights on private online platforms. But it would require considerable day-to-day government involvement. The need to reconcile the freedom of speech with the practical realities of online censorship would force policymakers to frame speech freedoms primarily in administrative rather than deregulatory terms—a nearly Copernican conceptual shift. Entrusting the power over content moderation to either the FCC or a new agency would create a new, immensely important quasi-constitutional institution, and it may therefore seem more appealing and familiar to lodge those powers in the judiciary instead. But having judges oversee content moderation standards will not change the novel and administrative character of the job. Eventually *someone* will be tasked with safety-checking the censorship algorithms, and doing so will require a kind of close, ongoing regulatory intervention in the public sphere that has no clear historical parallel.

Another disconcerting aspect of mandatory limits is less concrete but equally clear: mandatory limits have the potential to slowly dislodge the practical

127. *See* Frank Fagan, *Big Data Legal Scholarship: Toward a Research Program and Practitioner's Guide*, 20 VA. J.L. & TECH. 1, 59 (2016) ("The mechanics behind classification are straightforward. First, the analyst randomly samples a portion of her dataset and classifies the sample by hand. Second, the hand-classified sample trains the algorithm. Finally, the trained algorithm classifies the remaining documents of the dataset.").

operation of free speech principles from the First Amendment tradition. Those responsible for ensuring the quality of platforms' content moderation might be less concerned with the structure of free speech doctrine than with calibrating algorithmic error tolerances. And their decisions, unlike the often piquant and memorable First Amendment decisions of the Supreme Court, would contribute little to public consciousness of free speech issues. *Tinker v. Des Moines Independent Community School District* is a teachable case because of its memorable characters, imagery, and rhetoric.[128] Processes of regulatory approval have none of that literary or cultural cachet. This problem that is not just nostalgic or sentimental: later generations may well disengage from the civic value of free speech as its workings become more technocratic and inaccessible. But that sense of drift might be unavoidable—free speech in content moderation is bound to be an obscure business, whether the government oversees it or not.

Though mandatory limits are in some ways a queasy solution to the problem of online censorship, their central virtue should not be overlooked: namely, that they would subordinate private censorship to public law. The law would accomplish this purpose at a high price: a disturbing and unfamiliar expansion of governmental power into the private sphere. Yet this extension of law over the world of content moderation may establish norms that *constrain* the government's power to censor over the long run. After all, the techniques used by social media platforms to regulate speech today will inevitably influence governmental conduct tomorrow. As others have noted, it is a myth that the Internet is some wild and unregulable thing; the government has at least the *technical* capacity to exercise the same degree of control that Facebook does, only over the Internet as a whole.[129] The censorship questions raised in private settings today will eventually arise in public settings.

National security is the most likely stage for these efforts. President Obama leaned on YouTube to take down a provocative anti-Muslim video shortly after the Benghazi attacks,[130] and under the shield of the state action doctrine, the White House avoided falling afoul of *Brandenburg v. Ohio*'s test for incitement of imminent lawless action.[131] But other presidents may know or care too little to navigate these constitutional waters; instead, they may simply "close parts of the Internet."[132] In some cases, such an action may even be defensible. For example, the government may, in the future, restrain the online distribution of 3D-printable

128. 393 U.S. 503, 506 (1969) (affirming high schoolers' right to protest the Vietnam War by wearing black arm bands and noting that "[i]t can hardly be argued that either students or teachers shed their constitutional rights to freedom of speech or expression at the schoolhouse gate").

129. *See generally* GOLDSMITH & WU, *supra* note 40; E.H., supra note 40.

130. *See* Dawn C. Chmielewski, *'Innocence of Muslims' : Administration Asks YouTube to Review Video*, L.A. TIMES (Sept. 13, 2012), http://articles.latimes.com/2012/sep/13/entertainment/la-et-ct-administration-asks-youtube-to-review-innocence-of-muslims-video-20120913 [https://perma.cc/H3Q5-3H4C]. To its credit, YouTube refused. *See id.*

131. 395 U.S. 444, 447 (1969).

132. *See* Sam Frizell, *Donald Trump Wants to Close Off Parts of the Internet*, TIME (Dec. 16, 2015), http://time.com/4150891/republican-debate-donald-trump-internet/ [https://perma.cc/EQF9-3H7M].

dangerous objects,[133] or emails connected to ransomware attacks and other cybersecurity threats.

Once the government starts down this road, it will find itself engaging with problems and using techniques that are today being pioneered on social media. If there is already a precedent of social media applying the same techniques with impunity to more mundane speech issues, such as obscenity and true threats, then government actors are more likely to be tempted to carry over the most extraordinary types of restraints from the cyber and national security settings to other, more conventional speech concerns.

III. MANDATORY USER TOGGLES

A second broad approach to regulating content moderation—one that would enable "personal responsibility" from users rather than imposing central governmental oversight—would require online platforms to make content moderation settings togglable by users. Such an approach would be a spiritual successor to the "V-chip"—a user-programmable content filter that is federally required for all TV sets sold in the United States.[134]

The major search engines already allow users to toggle "adult" or "family" settings—a rational choice from a profit maximization standpoint.[135] Mark Zuckerberg's February 2017 manifesto, *Building Global Community*, suggested that Facebook may soon do the same.[136] But there is no reason to expect that platform owners' self-interests will always align with libertarian content moderation policies.

Mandatory toggling would present a few slim advantages over across-the-board limits. First, it may be more politically salable. A bill to extend speech rights online would tend to expose its supporters to a lot of risk, given the menagerie of unsavory speakers who would benefit from it. But if that extension of speech rights were paired with a guarantee that empowered the upstanding median voter to *shut off* the objectionable speech, then the overall boon to wholesomeness may offset the perception that the bill provided aid to the sleazy.

Beyond the politics, a toggling approach would avoid most, though not all, concerns about platform owners' speech and associational rights. Users could be informed that they are responsible for their own content moderation preferences,

---

133. The United States Department of State has already feebly attempted to restrain the online distribution of 3D-printable handguns, provoking a First Amendment challenge. *See* Def. Distributed v. U.S. Dep't of State, 838 F.3d 451, 454–56 (5th Cir. 2016); *see generally* Langvardt, *The Replicator*, *supra* note 49 at 101-110 (anticipating an eventual need to regulate a wider range of dangerous 3D-printable products by regulating the distribution of digital blueprints online).

134. *See* Telecommunications Act of 1996, Pub. L. No. 104-104, § 551(c), 110 Stat. 56, 141 (1996).

135. *See Block Explicit Results on Google Using SafeSearch*, GOOGLE, https://support.google.com/websearch/answer/510?co=GENIE.Platform%3DDesktop&hl=en [https://perma.cc/WJ55-92YJ] (last visited Apr. 10, 2018).

136. Mark Zuckerberg, *Building Global Community*, FACEBOOK (Feb. 16, 2017), https://www.facebook.com/notes/mark-zuckerberg/building-global-community/10154544292806634/ [https://perma.cc/2CDA-4QMA].

and that those preferences do not reflect the company's values. Companies might even recommend some slate of content moderation preferences as the "house blend" that best represents the corporate culture. To whatever extent diversity among corporate content moderation regimes is a free speech benefit—the point is, shall we say, abstract—the toggling approach would be an improvement over the mandatory limits approach.

Finally, a toggling approach would at least appear to relieve the government of the burden of designing the speech protections described in the previous section. If control over content moderation standards is punted to users, then the difficult questions of free speech principles can be deferred to some later case in which the government itself acts as the content moderator. These questions can then be decided by courts under First Amendment principles rather than a low-profile administrative body acting under an enabling statute.

In practice, however, a toggling approach could ultimately prove nearly as difficult to administer as the mandatory limits approach. One problem is relatively minor, and could be avoided by watering down the law. The range of censorship decisions must somehow be boiled down to a set of comprehensible, workable toggles: one for adult language, one for moderate displays of violence, one for extreme displays of violence, and so on. If these toggles are to give users meaningful choices, then a toggling law must guarantee that the underlying censorship tools being toggled are functional. This guarantee, of course, could be made stronger or weaker; but a rigorous guarantee would potentially embroil the government in many of the same administrative oversight difficulties as the mandatory limits approach discussed above.[137]

But there is also a major problem, which, unlike the minor problem, cannot be avoided by watering down the statute: which moderation settings will be togglable, and which will not be? At least some content moderation policies, such as the blocking of child pornography and other illegal material, should not be togglable—and it is inconceivable that Congress would ever require them to be.

Any regime of mandatory user toggles, then, would realistically be subject to at least some exceptions. Platforms would be permitted to adopt whatever content controls they like, so long as those controls are togglable. But they would be exempt from applying the user toggle to certain categories of unprotected speech, and that exemption from toggling would likely extend to other categories of unprotected speech as well—defamation, perhaps. And as a political reality, the exemption would probably also cover at least some constitutionally *protected* speech, including recruitment materials for international terrorist organizations, "fake news" from Russia, and so on—though these content-discriminatory extensions of the exemption would likely invite promising First Amendment challenges.[138]

137. *See supra* Section II.C.

138. Of course, it is also possible that Congress would rather let the material remain available and surveil those who read it.

At any rate, problems would arise in cases in which the platforms misjudged the boundaries of the exemptions. Defining the exemptions, interpreting them, and determining how to enforce them would require the government and the platform operators to assume essentially all of the conceptual and technical burdens that were involved in the mandatory limit-based approaches discussed above.[139] The difficulty involved in policing these lines means that a mandatory user toggle would be little more elegant or efficient than a mandatory limits approach.

Setting aside these administrative difficulties, there are broader questions about whether user toggles are adequate to restore what online censorship has placed at risk. User toggles would restore to users the right to see materials that might otherwise have been withheld from them, but they would not help the cause of confrontational speakers who wish to reach unwilling or reluctant beholders.

To a great extent, this is a good thing. Users of social media who face aggressive trolling campaigns, for instance, present an especially strong and sympathetic case for content filtration.[140] Other filtration options, however—options to shut out profanity, sexuality, ideological provocation, and so on—may, in the aggregate, frustrate speakers while stultifying listeners.

Viewers have always had the right to "avert[] their eyes," to borrow Justice Harlan's phrase in *Cohen v. California*.[141] But they have not until recently had the ability to avert them so completely. The genius of *Cohen*'s "avert your eyes" regime is that it does not really work: you avert your eyes only once you have already seen what you did not want to see. Some exposure to unwanted content is an inoculant, allowing the viewer to find the range of taste and opinion and to gauge her distance from the social periphery. It implants doubt and encourages mental toughness.

The alternative to *Cohen*'s approach appears in a lesser opinion written seven years later. In *FCC v. Pacifica Foundation*, Justice Stevens upheld the FCC's right to censor daytime broadcasts for foul language on the theory that inadvertent exposure to it "could have enlarged a child's vocabulary in an instant."[142] That kind of epistemic closure is arguably appropriate for children, but for adults—and ultimately for society—it could be infantilizing.

Yet one would be hard-pressed to argue that users should not have at least some ability to shield themselves from offensive or personally traumatic content. The most significant question in a toggling scheme may therefore have to do with the default settings. Turning off all content moderation by default would spark a nationwide freak-out until users figured out how to tweak their settings.

139. *See supra* Part II.

140. *See* Arthur Gaus, *Trolling Attacks and the Need for New Approaches to Privacy Torts*, 47 U.S.F. L. REV. 353, 356–60 (2012).

141. 403 U.S. 15, 21 (1971) (advising that those exposed to vulgar language printed on criminal defendant's jacket "could effectively avoid further bombardment of their sensibilities simply by averting their eyes").

142. 438 U.S. 726, 749 (1978).

Defaulting to an aggressive content moderation setting, on the other hand, would bias nationwide usage toward a childlike, highly censored speech environment.

In his manifesto, Mark Zuckerberg proposes a third way: that a user's default content controls, in the absence of any expressed choice, be set to those of the majority of users in the user's geographical area, "like a referendum."[143] The idea is reasonable, but it carries at least the potential of compounding Facebook's already polarizing "filter bubble"[144] with the well-documented tendency of Americans to segregate themselves into culturally homogeneous zip codes.[145] It seems fair to assume, at any rate, that many Americans would not touch the content settings at all. Despite great public fanfare at its introduction, few American parents have ever used the V-chip to moderate their children's television viewing.[146]

IV. MANDATORY DISCLOSURE

A lighter touch still would require the platforms to disclose information about their content moderation activities to the public. Such disclosures might help to dispel illusions that activity occurring on the platform is unmediated and neutral. It took an internal leak in 2016 to reveal that Facebook's "Trending Topics" sidebar was not purely algorithmic, but subject to human editorial control.[147] Until a second leak in 2017, the public had never seen Facebook's content moderation guidelines.[148] Regular disclosures, in theory, might help the public hold social platforms accountable for arbitrary or biased policies.[149]

Disclosure of algorithms, in particular, could facilitate what Maayan Perel and Niva Elkin-Koren call "black box tinkering"—that is, probing an algorithm's workings by testing it against litmus-test data sets.[150] It is already possible to tinker with content moderation algorithms today, but only if the tinkerer is willing to risk posting sensitive or unlawful material in a public forum and "testing" the

143. *See* Zuckerberg, *supra* note 136.

144. *See* CASS R. SUNSTEIN, REPUBLIC.COM 2.0 1–19 (1st ed. 2007) (envisioning, before modern social media took flight, the concepts of what we now call "filter bubbles" and "the Daily Me").

145. *See generally* BILL BISHOP, THE BIG SORT: WHY THE CLUSTERING OF LIKE-MINDED AMERICA IS TEARING US APART (2008) (reviewing Americans' geographic self-sorting along cultural, religious, and ideological lines since the 1970s).

146. Jim Rutenberg, *Survey Shows Few Parents Use TV V-Chip to Limit Children's Viewing*, N.Y. TIMES (Jul. 25, 2001), http://www.nytimes.com/2001/07/25/arts/survey-shows-few-parents-use-tv-v-chip-to-limit-children-s-viewing.html [https://perma.cc/Q8ZB-HVDE].

147. *See* Brian Feldman, *4 Takeaways From Facebook's Trending-Topics Controversy*, N.Y. MAG. (May 13, 2016, 5:24 PM), http://nymag.com/selectall/2016/05/four-takeaways-from-facebooks-trending-topics-controversy.html [https://perma.cc/8K2U-6C98].

148. *See* Jamie Grierson, *'No Grey Areas': Experts Urge Facebook to Change Moderation Policies*, THE GUARDIAN (May 22, 2017, 10:04 AM), https://www.theguardian.com/news/2017/may/22/no-grey-areas-experts-urge-facebook-to-change-moderation-policies [https://perma.cc/DY2Z-NXBN].

149. *See id.* ("These companies are hugely powerful and influential. They have given people a platform to do amazing and wonderful things but also dangerous and harmful things. Given the impact of the content decisions they make, their standards should be transparent and debated publicly, not decided behind closed doors." (quoting UK MP Yvette Cooper)).

150. *See* Perel & Elkin-Koren, *supra* note 126, at 185.

content moderation's ability to block that content. The danger here is that tinkerers must play with live ammunition: by attempting to post unlawful content, tinkerers risk violating the law.[151] Public disclosure of content moderation protocols, accompanied by a safe harbor for black box tinkering, could enable researchers to investigate the quality of content moderation algorithms in a laboratory setting.[152]

One likely constitutional objection is clear: the platforms will argue that content moderation algorithms comprise speech in the form of computer code, and mandatory disclosure of those algorithms would therefore be a form of compelled speech.[153] Apple has already made this argument in the aftermath of the San Bernadino mass shooting, after the Justice Department demanded that Apple supply law enforcement with a backdoor to the iPhone's encryption software.[154]

More seriously, platforms may be concerned about losing trade secrets or providing adversaries with road maps for evasion. These concerns might be allayed somewhat if certain limiting conditions are placed on disclosure. Presumably this problem could be overcome if the mandatory public disclosures were general rather than granular in nature, or if safeguards were put in place to prevent more specific disclosures from getting into the wrong hands. Disclosures might be made, for instance, only to trusted research institutions that comply with some defined registration and security certification process.

Even general, high level overviews of corporate content moderation policies could bring popular pressure to bear on the platforms. No one should be too sanguine, however, about market forces' ability to rein in private censorship, because the great online platforms are mostly insulated from market pressure. As discussed above, the platforms' strong network effects not only lock out competition, but makes competition undesirable.[155] Online citizens are in no position to "vote with their feet" by effecting a mass exodus from Facebook over censorship, and if they could, there is no reason to expect that they would. Consider the example of Gab, the Twitter alternative that proclaims it is "for creators who believe in free speech, individual liberty, and the free flow of information online."[156] Today it boasts only 170,000 users and is used almost exclusively by the extreme racist right.[157] An "alt" social network is about as enticing to the reasonable consumer as an unsubsidized high-risk insurance pool.

---

151. *See id.* at 212–17.

152. *See id.* at 217 (suggesting a safe harbor for *de minimis* legal violations occurring in the course of black box tinkering activity).

153. *See supra* Section II.A.

154. *See supra* note 50.

155. *See supra* Section II.C.1.

156. *Gab*, STARTENGINE, https://www.startengine.com/startup/gab [https://perma.cc/DG57-PNSJ] (last visited Apr. 10, 2018).

157. *See* Emma Grey Ellis, *Gab, the Alt-Right's Very Own Twitter, Is the Ultimate Filter Bubble*, WIRED (Sept. 14, 2016, 7:00 AM), https://www.wired.com/2016/09/gab-alt-rights-twitter-ultimate-filter-bubble/ [https://perma.cc/4G3J-JEAE]; Alina Selyukh, *Feeling Sidelined by Mainstream Social Media, Far-Right Users Jump to Gab*, NAT'L PUB. RADIO (May 21, 2017 6:46 AM), http://www.npr.org/sections/alltechconsidered/2017/05/21/529005840/feeling-sidelined-by-mainstream-social-media-far-right-users-jump-to-gab [https://perma.cc/W287-4ZZW].

And there is a deeper reason for doubt about the efficacy of disclosure. Even if we assume that market pressures and popular opinion will exercise strong influences on the platforms' censorship practices, why would we assume that the market and popular opinion would demand anything resembling the First Amendment? Is it not just as likely that the public would demand some nominal commitment to free speech, but with regular exceptions to accommodate majoritarian values? The whole point of placing the freedom of speech beyond the reach of democratic politics is, after all, to prevent censorship by popular demand.

V. LEAVING IT TO THE PRIVATE SECTOR

This leaves us with the fourth option: continuing to leave the entire business of content moderation to an unchecked private sector. This option represents the status quo, and for that reason some bias toward it is natural. But it is worth considering whether this fourth option represents a world we would choose *ex ante*.

On the plus side, corporate content moderators seem to be well-meaning and to take their jobs seriously.[158] Pitched debate takes place on social media platforms every day, and most users have never been personally censored. The system "works" reasonably well, and the owners of the platforms have incentives to ensure that it continues to work. But that everyday life goes on as usual does not imply that strong protections for speech are in place. The most likely reason that Facebook's content moderation policies are so broadly accepted is that most of the burden falls on marginal or unpopular speakers—exactly the speakers whom the law of free speech is traditionally concerned with protecting.

A content policy based on offensiveness aligns well with market incentives but poorly with the doctrine and priorities of First Amendment law. Facebook and Twitter, for instance, ban hate speech on the basis of racial, ethnic, or gender identity.[159] Maybe they are right to do so, even though that policy departs from the Supreme Court's approach. But suppose that all of the "big five" tech oligopolists—Apple, Amazon, Facebook, Microsoft, and Alphabet, Google's parent company[160]—adopted a joint policy to suppress hate speech wherever they can; and for the sake of the hypothetical, make Twitter a party to the agreement as well. Given the relative absence of alternative channels of communication, would it really be appropriate to leave such a complex and momentous social decision to the boards of six private corporations clustered in the San Francisco and Seattle metropolitan areas?

158. *See* Klonick, *supra* note 13, at 29–56 (providing a close, broadly sympathetic account of in-house content moderation practices and their evolution over time).

159. *See Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards#hate-speech [https://perma.cc/RMS6-PNC4] (last visited Aug. 2, 2017); *Hateful Conduct Policy*, TWITTER https://support.twitter.com/articles/20175050# [https://perma.cc/Z6KD-ZTQS].

160. *See* Farhad Manjoo, *Tech's Frightful Five: They've Got Us*, N.Y. TIMES (May 10, 2017), https://www.nytimes.com/2017/05/10/technology/techs-frightful-five-theyve-got-us.html [https://nyti.ms/2pwtHtt] (referring to the five tech giants as the "the handful of American technology companies that now dominate much of the global economy").

Some might say that the government should take a "wait and see" approach—hold off for now, as long as responsible people run the platforms, and step in later if their content moderation practices become abusive. But even supposing that the federal government will always have the final say in these matters—which is to assume that the government's efforts to liberalize the platforms would not themselves be held unconstitutional—it will be more difficult politically to take action *in response* to excessive censorship than to preempt it.

In a time of national emergency, for instance, the platforms may move to suppress the expression of certain despised viewpoints on their services. This may be a public relations maneuver, or it may be in response to informal pressure from government actors. The move may be justified as an effort to protect users from trauma, to promote public safety, or some similar goal—or it may be "an arbitrary decision" based on personal caprice.[161] Once such action has been taken, speakers who have been censored hold an extremely weak position. They have no existing constitutional or legal protections to draw on, and they are even more poorly positioned than usual to petition the political branches to create new legal protections.

Even to the extent that it is not formally recognized in law, the freedom of speech remains a prestigious ideological concept to which the unpopular can sometimes appeal as a source of political protection. But such tactics only work so long as the background ideological value of free speech remains robust. There is a danger that, after years of acquaintance with content moderation norms on social media, the public will become so inured to them that appeals to free speech norms will not resonate as deeply. Ultimately, then, the choice to defer to the platforms within their own spheres may have the spillover effects of diluting free speech as a public value and of inhibiting the government's ability to protect free speech through politics.

The laissez-faire approach is the most politically realistic one and is, on a certain view, the one that allows the marketplace of ideas to function as it should. But it leads to a disturbing outcome in which a small number of oligarchs take over responsibility for designing and implementing the system of free speech where it matters most. The nature of this power is distinct from any that has been exercised in the past by any private entity—save, arguably, the owners of company towns.[162] But company towns were always the exception, rather than the rule, and their ownership was far less concentrated.

Some may say, initially, that this is not such a bad outcome so long as the state is not involved. Yet the laissez-faire approach, counterintuitively, is the one that invites the worst abuses by the state. Here is a system with unprecedented censorship capabilities at a technical level, and because it is not managed by state actors, it is free from any clear constitutional limitation. It mediates a dominant and growing share of all online communication, and its private owners are few enough in number to operate as convenient "choke points" under pressure.

161. *See supra* note 20.
162. *See supra* Section II.B.1.

Governments, and especially executive officials, have every incentive to cultivate the "cooperation" of those who operate this system in censoring dangerous content.[163] And to make matters worse, these government actors must be careful, for legal reasons, to secure that cooperation on a strictly informal basis—that is, at the furthest distance possible from ordinary congressional or administrative procedure.

Under such a system, the shape of free speech will be determined by popular opinion, market pressures, governmental pressures, and managerial conscience. That is an extremely uncertain foundation for the future of free speech in "the most important place[] . . . for the exchange of views."[164]

CONCLUSION

The censorship capabilities of the Internet's moderators present us with a slow moving but immensely significant crisis for the future of free speech. The problem today is barely noticeable, and the judiciary still decides the central free speech issues of the day. But over time, both the First Amendment and the courts are on track to become more peripheral in the circle of free speech issues, and content moderators' technical powers rest like a "loaded gun" to be used under pressure from overreaching government officials.

The lawyers who oversee content moderation at the major social media platforms are in an impossible position. They cannot simply mandate that the platforms adopt First Amendment doctrine wholesale; it would not work operationally. Instead, in Jeffrey Rosen's words, they "are trying, in the face of great commercial pressures to the contrary, to enforce as much of the American free speech standards as possible."[165]

But entrusting the online freedom of speech to a small cadre of low-profile attorneys—whatever their skill and their good faith—can, in the long term, only corrode a transcendent public value. Consider, as one example, Facebook's policy on "Adult Nudity and Sexual Activity":

> We remove photographs of people displaying genitals or focusing in on fully exposed buttocks. We also restrict some images of female breasts if they include the nipple, but our intent is to allow images that are shared for medical or health purposes. We also allow photos of women actively engaged in breast-feeding or showing breasts with post-mastectomy scarring. We also allow photographs of paintings, sculptures, and other art that depicts nude figures. Restrictions on the display of sexual activity also apply to digitally created content unless the content is posted for educational, humorous, or satirical

163. After an attack on a diplomatic compound in Benghazi, Libya, the Obama Administration asked YouTube to "review" an inflammatory video thought to have provoked the attack. *See supra* note 130 and accompanying text.

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

165. Jeffrey Rosen, 2016 Richard S. Salant Lecture on Freedom of the Press at the Harvard Kennedy School's Shorenstein Center (Oct. 13, 2016).

> purposes. Explicit images of sexual intercourse are prohibited. Descriptions of sexual acts that go into vivid detail may also be removed.[166]

The point here is not to contrast the substance of Facebook's rules with those of the Supreme Court; some deviations there are only reasonable. The problem, instead, is the *slightness* of the doctrine and of the institution that has produced it. These are rules that would befit a small discussion group. As part of a plan to govern a society under a framework of civil liberties, however, they inspire no confidence. They are drafted on an ad hoc basis. Their author has a discernable point of view on specific contemporary controversies. They promise absolutely nothing: they may be rewritten tomorrow, and they may already have been rewritten behind closed doors. And the entire enterprise of writing these policies takes a back seat to the profit motive.

If you are comfortable with this approach, and you have faith that the well-meaning, blandly progressive oligopolists of the West Coast can secure the future of online free speech, ask yourself how you might feel if they were owned by someone with a different political or cultural baseline—the Walton family, or the Koch brothers, or the Breitbart-affiliated hedge-fund billionaire Robert Mercer.[167] And whoever is at the helm, how much faith do you have in the major online platforms to protect robust speech rights online during the next major national security crisis? It will be a matter of first impression, after all—remember that on September 11, 2001, not even Friendster, the proto-proto-Facebook, had yet come online.[168]

Individual speech rights on the Internet are important enough to deserve a legal charter. The range of options is unappealing, but one way or another, our society will make policy on this issue. By doing nothing, we already are.

166. *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards#nudity [https://perma.cc/A43B-VB62] (last visited Apr. 8, 2018).

167. *See* Carole Cadwalladr, *Robert Mercer: The Big Data Billionaire Waging War on Mainstream Media*, GUARDIAN (Feb. 26, 2017, 4:00 AM), https://www.theguardian.com/politics/2017/feb/26/robert-mercer-breitbart-war-on-media-steve-bannon-donald-trump-nigel-farage [https://perma.cc/3572-CG5M].

168. *See* Monica Riese, *The Definitive History of Social Media*, THE DAILY DOT (Sept. 12, 2016, 1:00 AM), https://www.dailydot.com/debug/history-of-social-media/ [https://perma.cc/HEJ5-WVYC].