Volume 67 | Issue 3 Article 1

9-9-2022

Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance

Brenda Dvoskin

VILLANOVA LAW REVIEW

VOLUME 67 2022 NUMBER 3

ARTICLES

REPRESENTATION WITHOUT ELECTIONS: CIVIL SOCIETY PARTICIPATION AS A REMEDY FOR THE DEMOCRATIC DEFICITS OF ONLINE SPEECH GOVERNANCE

BRENDA DVOSKIN*

ABSTRACT

Giant social media companies wield oversized power over what gets communicated online. Yet, it remains unclear how to hold that power accountable to the public. In democracies, governments are limited in how much they can regulate speech directly, creating an obstacle for efforts to make online speech governance more democratic. Corporations are touting civil society consultations to try to regain trust in how they write their rules. Scholars and lawmakers, too, are looking to mandate increased reliance on civil society participation as a means to democratize private governance without involving governmental bodies.

These proposals often make two assumptions. The first is that civil society participation operates outside the law, within the space that the First Amendment and Section 230 of the Communication Decency Act have carved out from the reach of courts and legislatures. The other is that if diverse civil society actors are involved, they will bring diverse viewpoints to bear in the shaping of policy outcomes.

* Doctoral Candidate, Harvard Law School; Affiliate, Berkman Klein Center for Internet & Society. I am deeply grateful to Yochai Benkler, Evelyn Douek, Noah Feldman, Janet Halley, David Kennedy, Gali Racabi, and Susan Silbey for their comments to prior drafts and multiple exchanges that informed this piece. This Article benefited from the suggestions of participants at the STS workshop at Harvard Law School, the Junior Scholars workshop at Cornell Law School, the Against Platform Determinism workshop organized by McGill Center for Media, Technology & Democracy and Data & Society, seminars at Universidad Di Tella and Universidad de Palermo, and conferences hosted by Law and Society and the Platform Governance Research Network. I am thankful to the *Villanova Law Review* team for their work on this piece. Most of all, I am thankful to the over sixty people who agreed to share their experiences and struggles in, with, and for online speech governance. This Article reflects developments through August 2021, when it was finalized for publication.

Drawing on over sixty original interviews with civil society actors and company representatives, this Article shows that both prevailing assumptions about civil society participation in content moderation are wrong. First, the governance of online speech has been anything but lawless. Law has played a major role in constituting companies' and civil society's interests. Not only have advocates gravitated towards the legal frameworks that advance their interests, but available legal texts have constrained and expanded their imagination to formulate normative demands.

Second, civil society participation has had distributive effects among advocacy groups. Those who favor more restrictive online speech rules have been more effective advocates in the last few years than those who believe that restrictions on speech undermine freedom of expression. This Article casts doubt on the claim that this trend responds to a shift in social preferences over how to regulate freedom of expression. The analysis suggests instead that advocates interested in advancing more aggressive speech regulations have been more successful than their counterparts at mobilizing legal discourses and solidifying a lobbying agenda. The Article concludes with proposals to adjust and strengthen civil society participation as a democratizing tool.

CONTENTS

INTRODUCTION 450
I. THE PROJECT: CIVIL SOCIETY ORGANIZATIONS AS REPRESENTATIVES OF THE PUBLIC INTEREST 455
II. HOW IT HAS WORKED: THE EVOLUTION OF CONTENT MODERATION 462
A. *Outcomes* 462
B. *Winners and Losers* 465
C. *Possible Explanations* 468
1. *Does Civil Society Participation Matter?* 468
2. *Public Choice Theory* 473
3. *A Technology-Driven Shift in Social Norms* 476
III. THE ASYMMETRIC U.S. CIVIL SOCIETY 478
A. *Speech Regulationists* 478
1. *Civil Rights* 479
2. *Silencing Speech* 482
3. *"Free Speech Is Not the Same as Free Reach"* 485
B. *Speech Protectionists* 487
1. *The Modern Public Square* 489
2. *The Administrative Law Project* 492
3. *Human Rights* 495
IV. THE EFFECTS OF LAW IN ONLINE SPEECH GOVERNANCE 499
V. CIVIL SOCIETY PARTICIPATION AS A DEMOCRATIZING MECHANISM 502
CONCLUSION 506

INTRODUCTION

SCHOLARS have described social media companies as the new governors,[1] deciders,[2] sovereigns,[3] and regulators[4] of freedom of expression on the internet. These labels reflect not only a broad consensus about the fact that these companies exercise oversized power,[5] but also that these companies are carrying out a task—the regulation of speech—that one may imagine has traditionally been the prerogative of governmental bodies.[6] This conceptualization of corporate actions has had deep normative consequences; if companies are substituting the role of governments, their internal rules ought to be accountable to the public interest in a way similar as governments' rules.

Where citizens cannot vote on corporate behavior, how should social media companies align their actions with the public interest? If the concern is that corporations are taking over governments' prerogatives, the most immediate remedy would be for democratic institutions—such as legislatures or administrative agencies—to replace them. In this sense, governments can set some parameters for content moderation.[7] In the

1. *See* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1603 (2018).

2. *See generally* Jeffrey Rosen, *The Deciders: Facebook, Google, and the Future of Privacy and Free Speech*, BROOKINGS (May 2, 2011), https://www.brookings.edu/research/the-deciders-facebook-google-and-the-future-of-privacy-and-free-speech [https://perma.cc/MGE6-NJ2A].

3. *See generally* Jonathan Peters, *The "Sovereigns of Cyberspace" and State Action: The First Amendment's Application—or Lack Thereof—to Third-Party Platforms*, 32 BERKELEY TECH. L.J. 989 (2018).

4. Susan Benesch & Rebecca MacKinnon, *The Innocence of YouTube*, FOREIGN POL'Y (Oct. 5, 2012, 4:47 PM), http://foreignpolicy.com/2012/10/05/the-innocence-of-youtube/ [https://perma.cc/C5AK-GBRS].

5. Yochai Benkler, *Degrees of Freedom, Dimensions of Power*, 145 DAEDALUS 18 (2016); YOCHAI BENKLER, ROBERT FARIS & HAL ROBERTS, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION, AND RADICALIZATION IN AMERICAN POLITICS 4 (2018); Jack Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2021 (2018).

6. *See, e.g.*, Julie Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 199 (2017) ("Dominant platforms' role in the international legal order increasingly resembles that of sovereign states."); JILLIAN YORK, SILICON VALUES: THE FUTURE OF FREE SPEECH UNDER SURVEILLANCE CAPITALISM 8 (2021) (describing how with emergence of "the nation-state and the lesser governance structures contained within it became the predominant arbiter of what people can do or say, and which information they can access" until the rise of the "new gatekeepers," social media companies); Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, HOOVER INST. AEGIS PAPER SERIES, 2–3 (2019) ("[P]latforms can take on and displace traditional state functions, operating the modern equivalent of the public square or the post office, without assuming state responsibilities[.]"); Richard Ashby Wilson & Molly Land, *Hate Speech on Social Media: Content Moderation in Context*, 52 CONN. L. REV. 1029, 1032 (2020) ("Governments are no longer the primary regulators of speech. Their regulatory capacity has been so far outstripped by some of the largest companies in the world . . . ." (footnotes omitted)).

7. *See, e.g.*, Netzwerkdurchsetzungsgesetz [NetzDG] [Network Enforcement Act], Oct. 1, 2017, https://www.bmj.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf?__blob=publicationFile&v=2 [https://perma.cc/

United States, proposals both to amend Section 230 of the Communications Decency Act[8] and to introduce must-carry obligations would restrict the current degree of freedom that corporations enjoy when deciding how to moderate their platforms.[9] However, governments—and especially the U.S. government—are ill-situated to be the ones drafting specific content moderation rules.

The limitations on government regulation of online speech are technical, legal, and normative. Technically, government intervention is limited by potential disconnect between granular, rapidly evolving content moderation rules and democratic rule-making processes.[10] Legally, current First Amendment jurisprudence prevents the U.S. government from mandating companies either to host content or to eliminate a big part of the controversial content. Because of the First Amendment's state action requirement, a company is not required to leave users' speech up, and most of the speech that companies regulate is protected under the First Amendment. Thus, even if companies were to bear responsibility for hosting illegal content generated by their users, they would not be responsible for most of the "problematic" speech that is constitutionally protected.[11] More fundamentally, most scholars and large portions of the public expect companies to adopt rules that differ from state regulation.[12]

This scenario poses the most challenging dilemma in online speech governance; it is problematic that private actors are overtaking government functions, but governments should constrain themselves from carrying out those functions. The administrative law project offers a possible way out.[13] The project borrows administrative law principles such as trans-

---

C4DA-CJ2P] (Ger.); *Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and Amending Directive 2000/31/EC*, art. 2, COM, (2020) 825 final (Dec. 15, 2020) [hereinafter *Digital Service Act*].

8. 47 U.S.C. § 230 (1996).

9. *See* Zoe Bedell & John Major, *What's Next for Section 230? A Roundup of Proposals*, LAWFARE (July 29, 2020, 9:01 AM), https://www.lawfareblog.com/whats-next-section-230-roundup-proposals [https://perma.cc/4T8Y-8JAC]; *see generally* Genevieve Lakier, *The Non-First Amendment Law of Freedom of Speech*, 134 HARV. L. REV. 2299 (2021) (showing that imposing duties on social media to carry some user-generated content would not necessarily violate the First Amendment).

10. Monika Bickert, *Defining the Boundaries of Free Speech on Social Media*, *in* THE FREE SPEECH CENTURY, 254, 257 (Lee C. Bollinger & Geoffrey Stone eds., 2018) ("[T]ransferring that role of content standard management entirely to governments or external bodies is not just impractical; it is impossible to do in a way that does not require the external body to become a smaller version of the same social media company[.]"); Evelyn Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation*, HOOVER INST. AEGIS PAPER SERIES, 1, 7 (2019).

11. Keller, *supra* note 6, at 4.

12. *Id.* at 13; *see infra* Section II.C.3.

13. Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 71–78 (2019) (grouping the projects of transparency, due process, reasoned decision-making, and participation together

parency, due process, and public participation as a means to democratize or legitimize companies' exercise of power. Jonathan Zittrain and John Bowers refer to this new phase in the short history of content moderation as the process era: a time in which it is necessary to design procedures to reorganize and legitimize companies' power.[14]

The administrative law approach generally defines public participation as companies engaging civil society organizations as representatives of the public interest.[15] Under the Administrative Procedure Act,[16] federal agencies must solicit and consider the views of the public as part of their rulemaking processes. Supporters of these notice-and-comment procedures believe that they can help democratize administrative agencies that are not directly subjected to electoral politics.[17] Civil society's participation in content governance could have analogous democratizing effects. In fact, in the case of other global governance institutions, equivalent mechanisms for civil society participation are often proposed with the same goal.[18]

However, the enthusiasm for broader civil society participation often overlooks the literature analyzing the problems of capture, imbalances of power, and barriers to participation in notice-and-comment procedures.[19]

---

under the administrative law umbrella); *see also* Evelyn Douek, *Content Moderation as Administration*, 136 HARV. L. REV. (forthcoming 2022).

14. John Bowers & Jonathan Zittrain, *Answering Impossible Questions: Content Governance in an Age of Disinformation*, 1 HARV. KENNEDY SCH. MISINFORMATION REV. 1, 4 (2020).

15. Bloch-Wehba, *supra* note 13, at 76 ("[P]latforms could adopt rigorous 'notice-and-comment' procedures for changes to their privacy policies, [Terms of Service (ToS)], and community standards[.]"). *See generally* TERRY MACDONALD, GLOBAL STAKEHOLDER DEMOCRACY: POWER AND REPRESENTATION BEYOND LIBERAL STATES (2008).

16. Administrative Procedure Act, 5 U.S.C. § 553 (1966).

17. Michael Asimow, *On Pressing McNollagst to the Limits: The Problem of Regulatory Costs*, 57 L. & CONTEMP. PROBS. 127, 129 (1994) ("[Notice-and-comments procedure] provides an ingenious substitute for the lack of electoral accountability of agency heads. Indeed, rulemaking procedures are refreshingly democratic: people who care about legislative outcomes produced by agencies have a structured opportunity to provide input into the decisionmaking process.").

18. Nahuel Maisley, *The International Right of Rights? Article 25(a) of the ICCPR as a Human Right to Take Part in International Law-Making*, 28 EUR. J. INT'L L. 89, 95 (2017). On the global administrative law project to address democratic deficits in international law making, see generally Benedict Kingsbury, Nico Krisch & Richard Stewart, *The Emergence of Global Administrative Law*, 68 L. & CONTEMP. PROBS. 15 (2005); *see also* Benedict Kingsbury, *The Administrative Law Frontier in Global Governance*, 99 PROC. AM. SOC'Y INT'L L. 143 (2005); Benedict Kingsbury & Nico Krisch, *Global Governance and Global Administrative Law in the International Legal Order*, 17 EUR. J. INT'L. L. 1 (2006). For further bibliographical resources, see *A Global Administrative Law Bibliography*, 68 L. & CONTEMP. PROBS. 357 (2005); *see also* David Kennedy, *The Mystery of Global Governance*, 34 OHIO N.U. L. REV. 827, 842 (2008) (situating this literature in the larger field of global governance).

19. *See, e.g.*, Jim Rossi, *Participation Run Amok: The Costs of Mass Participation for Deliberative Agency Decisionmaking*, 92 NW. U. L. REV. 173 (1997); *see also* Cary Coglianese, *Citizen Participation in Rulemaking: Past, Present, and Future*, 55 DUKE L.J. 943

Indeed, these projects often assume that civil society participation is a neutral vehicle useful to represent all viewpoints in the society.

This Article explores whether these governance devices can be effective in making private speech regulation accountable to the public interest. In answering this question, the Article presents two main findings. First, companies' engagement with civil society has not been neutral. Rather, the evolution of content moderation has favored the views of those advocates who support stricter restrictions on hate speech, harassment, violent and extremist content, bullying, and other forms of harmful speech. Although many would embrace these motives, this Article's findings show that civil society engagement, while apparently neutral, has involved a clear substantive orientation in practice.

The second finding illustrates how law has played a significant role in defining how civil society and private companies engage. At first glance, it seems like these conversations happen outside the law.[20] Section 230 of the Communications Decency Act protects platforms from most claims that could arise from their decision to host or delete content. Excluding a few exceptions, the law does not say much about which rules platforms should adopt. However, this Article argues that legal frameworks have had a strong impact not only on how civil society actors express their demands, but also on how they imagine their substantive agendas. Law is not only the language advocates use to articulate their interests; it also plays a significant role in constituting those interests.[21] Looking at the role the law plays in this apparently lawless space helps explain the relative power of different advocacy groups *vis-à-vis* companies.

Overall, this Article is an invitation for caution about the projects that aim at replicating state features within corporations. We would benefit from putting these projects' promise for substantive neutrality into question. In order to evaluate civil society participation, it is necessary to be aware of their distributional and normative implications.

This Article looks at governance mechanisms not from a hypothetical perspective, but from a perspective grounded in current models for allowing civil society to participate in content governance. The Article's findings are based on interviews with over sixty participants who include civil society actors that regularly engage with companies, as well as mem-

(2006); Jason Webb Yackee & Susan Webb Yackee, *A Bias Towards Business? Assessing Interest Group Influence on the U.S. Bureaucracy*, 68 J. POL. 128 (2006); *see infra* Part II. *See generally* BLAKE EMERSON, THE PUBLIC'S LAW: ORIGINS AND ARCHITECTURE OF PROGRESSIVE DEMOCRACY (2019).

20. *See* NICOLAS SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 9 (2019) (arguing that platforms govern in a lawless space).

21. *See* Amy Kapczynski, *The Access to Knowledge Mobilization and the New Politics of Intellectual Property*, 117 YALE L.J. 804, 860 (2008) (describing the influence of legal frameworks in the constitution of social movements' interests in the access to knowledge movement); *see also* JANET HALLEY, SPLIT DECISIONS: HOW AND WHY TO TAKE A BREAK FROM FEMINISM 43, 323 (2006) (describing how feminist movements have changed their aims through their engagement with legal frameworks).

bers of the Facebook Content Policy Stakeholder Engagement team and representatives of other companies. I conducted the interviews between April and November 2020 by videoconference due to the COVID-19 pandemic. The first set of interviews was conducted with thirty-seven individuals who are academics or advocates and who have engaged with major social media platforms in different formats to discuss their content moderation rules and enforcement.[22] This sample represents different types of actors: institutions with diverse political views and divergent opinions on how to balance freedom of expression with other values. The study also includes nine interviews with the members of the Facebook Content Policy Stakeholder Engagement team and two interviews with employees of other large companies.[23] I conducted the interviews with Facebook under an agreement. Per the agreement, Facebook reviewed a draft of this manuscript before publication for the purpose of identifying confidential information. Facebook did not request any removal of information. Seventeen additional interviews with advocates based in Latin America have provided supplementary context on how companies engage with external stakeholders.

The scope of this Article's analysis is limited to civil society in the United States. Selecting a specific jurisdiction makes it possible to study the role that legal frameworks play in shaping actors' interests and collective negotiations. To that end, no jurisdiction has been more influential than the United States: U.S. civil society has a long history of engaging with social media companies since they launched and has remained among the most consulted organizations globally.[24]

The Article is structured as follows. Part I explains why civil society participation is an attractive path for content governance. This Part maps the sheer number of calls for greater involvement of civil society as representatives of the public interest, as well as how civil society participation has taken place to date. Part II presents an empirical puzzle, as it describes the evolution of content moderation rules over the years and offers competing explanations to account for the relationship between those normative outcomes and the increasing participation of civil society. Part III argues that the way law has shaped advocates' interests and strategies

22. Due to the COVID-19 pandemic, I conducted all interviews over videoconference platform Zoom. I conducted all interviews with the informed consent of participants pursuant to a protocol approved by the Harvard Institutional Review Board. To protect confidentiality, all interviews cited in this article are referenced by interview number and date. All interview transcripts are on file with the author.

23. Facebook, now "Meta," was the only company that responded to my request to participate in the study. Employees at other tech companies agreed to be interviewed anonymously. While the interviews are helpful as general background, I do not mention any specific information that could be traced back to them in order to preserve their confidentiality.

24. Interviews with organizations and advocates based in Latin America show that they are more often consulted about the enforcement of policies in their regions and less significantly involved in policy development.

contributes to our understanding of the relative success or failure of different advocacy projects as well as the normative evolution of online speech governance. Part IV synthesizes the role of law in online speech governance. Part V offers some proposals to make civil society engagement a better mechanism for more democratic governance.

I. THE PROJECT: CIVIL SOCIETY ORGANIZATIONS AS REPRESENTATIVES OF THE PUBLIC INTEREST

Legal discussions about online speech often center around Section 230 of the Communications Decency Act.[25] This law immunizes companies from most legal claims users could bring based on corporate decisions regarding user-generated content.[26] Accordingly, it drastically limits the power of courts and state legislatures to interfere with companies' decisions regarding content moderation.

But more fundamental legal limitations on governmental action arise from the Constitution. Current First Amendment jurisprudence limits government's options in three ways. First, most of the speech that elicits conflict on social media is constitutionally protected. For example, adult pornography,[27] violent content that is not intended to provoke imminent illegal activity,[28] and falsehoods[29] are all protected by the First Amendment. Because the speech that sparks most controversies is legal, there is not much that the government can do to restrict it. Second, courts cannot force companies to allow user-generated content because social media corporations are not state actors; the First Amendment only protects individuals' right to be free from government interference of protected speech.[30] Finally, if the federal government decided to impose new duties on social media companies to host some types of speech, it would face significant challenges. Although not an impossible legislative option, any restriction would have to respect companies' own First Amendment rights to make editorial decisions.[31]

Civil society participation is the last best hope in many fields where states are not able to deliver on the promise of self-government. Salient

25. *See e.g.*, JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET (2019).

26. 47 U.S.C. § 230 (2012); *see also* Zeran v. Am. Online, Inc., 129 F.3d 327, 330 (4th Cir. 1997) (establishing the expansive interpretation of § 230 that continues today); Robert Bartlett, *Developments in the Law — The Law of Cyberspace*, 112 HARV. L. REV. 1574, 1613 (1999) (describing Zeran's holding as a "broad interpretation of § 230").

27. Am. Booksellers Ass'n v. Hudnut, 771 F.2d 323, 332–34 (7th Cir. 1985), *aff'd* 475 U.S. 1001 (1986).

28. Brandenburg v. Ohio, 395 U.S. 444, 448–49 (1969).

29. United States v. Alvarez, 567 U.S. 709, 728–30 (2012).

30. *See infra* Section III.B.1.

31. Keller, *supra* note 6, at 17.

examples are administrative agencies[32] and international institutions such as the European Union and the United Nations.[33] International organizations may seem too far removed from citizens' control, so citizens require other means of public participation through voluntary associations. Eduardo Szazi, discussing the legitimacy of civil society organizations to take part in internal lawmaking, poses the question: "if national parliaments are no longer capable of dealing with the complexities of contemporary globalized life and, in turn, the international sphere does not possess a parliament-like institution, how will society be able to cope with the challenges of ensuring pluralist democracy in the 21st century . . . ?"[34] In different fields that cross jurisdictional boundaries such as global finance, environmental law, and public international law, part of the response has been to include civil society organizations in governance structures.[35]

In content moderation discussions, civil society participation as a means to address the democratic deficits of corporate governance is ubiquitous not only in the academic literature, but also in initiatives from nonprofits, international organizations, and even in proposed legislation. Susan Benesch highlights "a growing consensus . . . that [online service providers] should not write and apply rules entirely on their own" and should find ways to engage users or those who can advocate on behalf of their interests.[36] In his book *Lawless*, Nicolas Suzor claims that, "[consultation with a wide range of civil society groups] is exactly the type of broad consultation that is expected of companies making decisions that have an impact on human rights[.]"[37] Kate Klonick as well as Ethan Zuckerman and Jillian York have argued that civil society organizations can play an

32. Asimow, *supra* note 17, at 129; Admin. Conf. of the U.S., Recommendation 2018-7, *Public Engagement in Rulemaking*, 84 Fed. Reg. 2139, 2146 (Feb. 6, 2019) ("Robust public participation is vital to the rulemaking process. By providing opportunities for public input and dialogue, agencies can obtain more comprehensive information, enhance the legitimacy and accountability of their decisions, and increase public support for their rules.").

33. *See* Maisley, *supra* note 18, at 104; *see also* Francesca E. Bignami, *The Democratic Deficit in the European Community Rulemaking: A Call for Notice and Comment in Comitology*, 40 HARV. INT'L L.J. 451, 452 (1999); Allen Buchanan & Robert O. Keohane, *The Legitimacy of Global Governance Institutions*, 20 ETHICS & INT'L AFF. 405, 432 (2006); Gráinne de Búrca, *The Quest for Legitimacy in the European Union*, 59 MOD. L. REV. 349, 356 (1996).

34. EDUARDO SZAZI, NGOS: LEGITIMATE SUBJECTS OF INTERNATIONAL LAW 271 (2012).

35. *See generally* CIVIL SOCIETY AND GLOBAL FINANCE (Jan Aart Scholte & Albrecht Schnabel eds., 2002); PARTICIPANTS IN THE INTERNATIONAL LEGAL SYSTEM: MULTIPLE PERSPECTIVES ON NON-STATE ACTORS IN INTERNATIONAL LAW (Jean d'Aspremont ed., 2011).

36. Susan Benesch, *Proposals for Improved Regulation of Harmful Online Content*, *in* REDUCING ONLINE HATE SPEECH: RECOMMENDATIONS FOR SOCIAL MEDIA COMPANIES AND INTERNET INTERMEDIARIES 274 (Yuval Shany ed., 2020).

37. SUZOR, *supra* note 20, at 126.

important role in pushing companies to change their rules.[38] Furthermore, Hannah Bloch-Wehba has proposed that companies adopt notice-and-comment procedures to give civil society a significant opportunity to participate in online speech governance.[39]

Overseas regulators have begun to design mandates to consult with civil society. For example, The Code of Practice on Disinformation is a set of best practices designed by the European Commission. Facebook, Google, Twitter, Mozilla, Microsoft, and TikTok have signed it.[40] The signatories commit to partner with civil society to support actions to combat disinformation.[41] Also, The European Commission's legislative proposal Digital Service Act (DSA) package would legally require companies operating in Europe to engage civil society organizations in different contexts as part of their due diligence obligations.[42] Finally, the DSA and the UK Online Safety Bill, both released in 2021, would mandate risk assessments, which would encourage engagement of experts and stakeholders.[43]

The Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online is an action plan led by the presidents of New Zealand and France, joined by other governments, and adopted by tech companies.[44] The initiative was formed following the massacre in Christchurch, New Zealand that was live-streamed on Facebook.[45] Pursuant to the action plan, governments and companies pledge to "work with civil society to promote community-led efforts" and to "recogni[z]e the important role of civil society in supporting work on the issues and commitments in the Call."[46] Along the same lines, the Global Internet Forum to Counter Terrorism, a consortium of companies collaborating to prevent terrorist and extremist content on digital platforms, created an Independent Advisory

38. Klonick, *supra* note 1, at 1655–57; Jillian York & Ethan Zuckerman, *Moderating the Public Sphere*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 137, 157 (Rikke Frank Jørgensen ed., 2019).

39. Bloch-Wehba, *supra* note 13, at 76, 80.

40. *Code of Practice on Disinformation*, EUROPEAN COMMISSION, https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation [https://perma.cc/NVF9-WKBJ] (last visited July 14, 2022).

41. *Id.*

42. *Digital Service Act*, *supra* note 7, arts. 35–37.

43. *Digital Service Act*, *supra* note 7, at Chapter III, Section 4; *Draft Online Safety Bill* 2021, c. 2 (UK), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/985033/Draft_Online_Safety_Bill_Bookmarked.pdf [https://perma.cc/VP5M-XC6B]; Special Representative of the Secretary-General on the issue of human rights and transnational corporations, *Human Rights Impact Assessments - Resolving Key Methodological Questions*, A/HRC/4/74 (Feb. 5, 2007) https://undocs.org/en/A/HRC/4/74 [permalink not available] ("Engagement of human rights experts and local stakeholders is critical" to conducting a human rights impact assessment).

44. *The Call*, CHRISTCHURCH CALL, https://www.christchurchcall.com/call.html [https://perma.cc/QWL2-LMJE] (last visited July 14, 2022).

45. *See id.*

46. *Id.*

Committee to engage nongovernmental organizations and other stakeholders.[47]

The Global Network Initiative (GNI) is an initiative joined by academics, civil society organizations, and tech companies such as Facebook, Google, and Microsoft.[48] Members of the initiative commit to implement the GNI Principles, including multi-stakeholder collaboration.[49] Ranking Digital Rights (RDR) is an initiative founded by Rebecca MacKinnon that produces an annual digital rights corporate accountability index that ranks companies according to how closely they follow a set of good practices.[50] Stakeholder engagement is one of the indicators the initiative measures.[51] To assess how well a company performs on this indicator, they ask whether the company takes part in one or more multi-stakeholder initiatives, whether the company "engages systematically and on a regular basis with non-industry and non-governmental stakeholders on freedom of expression and privacy issues," or whether the company "initiates or participates in meetings with stakeholders that represent, advocate on behalf of, or are people whose rights to freedom of expression and information and to privacy are directly impacted by the company's business."[52]

The Manila Principles on Intermediary Liability are a set of best practices endorsed by over a hundred organizations from around the world.[53] The principles aimed at guiding governments' and companies' practices state that accountability requires that intermediaries, civil society, and governments work together.[54] Article 19, an acclaimed worldwide organization, has proposed the creation of Social Media Councils. This project aims at creating a governance structure led by civil society organizations.[55]

---

47. *Independent Advisory Committee*, GIFCT, https://gifct.org/governance/#civil-society [https://perma.cc/G7FD-NDDS] (last visited July 14, 2021). *But see* Emma Llansó, *Human Rights NGOs in Coalition Letter to GIFCT*, CTR. FOR DEMOCRACY AND TECH. BLOG (July 30, 2020), https://cdt.org/insights/human-rights-ngos-in-coalition-letter-to-gifct/ [https://perma.cc/U7QD-YEHS] (highlighting that the Independent Advisory Committee is insufficient to engage adequately with civil society).

48. *The GNI Principles*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/gni-principles/ [https://perma.cc/N6LW-8LGX] (last visited July 14, 2022).

49. *Id.*

50. *Who We Are*, RANKING DIGIT. RTS., https://rankingdigitalrights.org/who-we-are/ [https://perma.cc/ZL9S-JADU] (last visited July 14, 2022).

51. *2020 Ranking Digital Rights Corporate Accountability Index, G5. Stakeholder engagement and accountability*, RANKING DIGIT. RTS., https://rankingdigitalrights.org/index2020/indicators/G5 [https://perma.cc/F4GQ-ZBEP] (last visited July 14, 2022).

52. *Id.*

53. MANILA PRINCIPLES ON INTERMEDIARY LIABILITY, https://manilaprinciples.org/principles.html [https://perma.cc/8QBA-AJVP] (last visited July 14, 2022).

54. *Id.* ("Transparency and accountability must be built into laws and content restriction policies and practices.").

55. *Social Media Councils*, ARTICLE 19, https://www.article19.org/social-media-councils/ [https://perma.cc/E3DC-8PTA] (last visited July 14, 2022).

Other well-known initiatives that highlight engagement with civil society as an essential component of internet governance include the Freedom Online Coalition[56] and the Internet & Jurisdiction Policy Network.[57] Other proposals have focused on partnerships with local organizations working on the ground.[58] Overall, engagement with civil society organizations is now considered good practice and an important feature of online speech governance.[59]

These calls are supported by United Nations agencies and documents. The United Nation's Guiding Principles on Business and Human Rights (UNGPs) provide that companies should consult those affected by their businesses either directly or by consulting experts, human rights defenders, and others from civil society.[60] Building on that instrument, the B-Tech project of the United Nation's Human Rights Committee has said on various occasions that tech companies should engage in partnerships with civil society "as a critical aspect of operating with respect for human rights."[61] The United Nations Educational, Scientific and Cultural Organization (UNESCO) "Social Media 4 Peace" project proposes the creation of a multi-stakeholder platform as a means to tackle harmful online con-

56. FREEDOM ONLINE COALITION, https://freedomonlinecoalition.com/ [https://perma.cc/RB7L-X4UW] (last visited May 29, 2022).

57. INTERNET & JURISDICTION POLICY NETWORK, https://www.internetjurisdiction.net/ [https://perma.cc/57KJ-G69T] (last visited July 14, 2022).

58. Center for American Progress, Color of Change, Free Press, Lawyer's Committee for Civil Rights Under Law, National Hispanic Media Coalition & Southern Poverty Law Center, *Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities*, CHANGE THE TERMS, https://assets.website-files.com/5bba6f4828dfc3686095bf6b/5bd0e36186e28d35874f0909_Recommended%20Internet%20Company%20Corporate%20Policies%20%20Terms%20of%20Service_final-10-24.pdf [https://perma.cc/TKD5-7L5H] (last visited July 14, 2022) [hereinafter *Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities*].

59. *See, e.g.*, Alison Taylor, Charlotte Bancilhon, Jonathan Morris & Cecile Oger, *Five-Step Approach to Stakeholder Engagement*, BUS. & SOC. RESP., (Apr. 29, 2019), https://www.bsr.org/en/our-insights/report-view/stakeholder-engagement-five-step-approach-toolkit?utm_source=communitystandardsblog&utm_medium=landingpg&utm_campaign=stakeholder-launch [https://perma.cc/B72G-Z6BL].

60. John Ruggie (Special Representative of the Secretary–General), *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter *UNGPs*], Comment to Principle 18.

61. B-Tech, *Addressing Business Model Related Human Rights Risks*, UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, https://www.ohchr.org/Documents/Issues/Business/B-Tech/B_Tech_Foundational_Paper.pdf [https://perma.cc/HHX3-ZW8W] (last visited May 29, 2022); *see also* B-Tech, *Designing and Implementing Effective Company-based Grievance Mechanisms*, UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, https://www.ohchr.org/Documents/Issues/Business/B-Tech/access-to-remedy-company-based-grievance-mechanisms.pdf [https://perma.cc/6J7H-AM2L] (last visited July 21, 2022); B-Tech, *Key Characteristics of Business Respect for Human Rights*, UNITED NATIONS HUM. RTS. OFF. HIGH COMM'R, https://www.ohchr.org/Documents/Issues/Business/B-Tech/key-characteristics-business-respect.pdf [https://perma.cc/4XX2-XXF5] (last visited July 21, 2022).

tent.[62] In his 2018 report, then U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, noted that rule-making and product development processes should incorporate input "from affected rights holders (or their representatives) and relevant local or subject matter experts" as an integral component of due diligence.[63]

Companies have embraced the consultation project and implemented various mechanisms to engage civil society. Policy announcements often state that they have been preceded by consultation with civil society.[64] In 2020, TikTok created a Content Advisory Council in the United States and, later, it added councils in other regions.[65] In May 2020, Twitch announced the creation of the Twitch Safety Advisory Council.[66] In December 2019, Twitter revamped its Trust and Safety Council, which was created in 2016.[67] The Facebook Oversight Board, launched in 2021, includes both formal appointments of civil society representatives as decision makers and channels for the public to submit comments in an

62. *Social Media 4 Peace*, UNESCO, https://en.unesco.org/social-media-4-peace [https://perma.cc/H7F9-QUJC] (last visited July 21, 2022).

63. David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, No. A/HCR/38/35 (Jun. 2018) [hereinafter *SR Report 2018*].

64. *See, e.g.*, *Standing Against Hate*, META (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate/ [https://perma.cc/YK9Y-JSGK]; Monika Bickert, *Enforcing Against Manipulated Media*, META (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/L373-4LLL]; Vanessa Pappas, *Combating Misinformation and Election Interference on TikTok*, TIKTOK (Aug. 5, 2020), https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok [https://perma.cc/2G46-88DR]; Twitter Safety, *Updating Our Rules Against Hateful Conduct*, TWITTER BLOG (July 9, 2019), https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html [https://perma.cc/AF6H-DYBB]; Vijaya Gadde & Del Harvey, *Creating New Policies Together*, TWITTER BLOG (Sept. 25, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Creating-new-policies-together.html [https://perma.cc/36L4-UY2C].

65. Vanessa Pappas, *Introducing TikTok Content Advisory Council*, TIKTOK (Mar. 18, 2020), https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council [https://perma.cc/47CK-KFKM]; Arjun Narayan Bettadopur, *Introducing the TikTok Asia Pacific Safety Advisory Council*, TIKTOK (Sept. 21, 2020), https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council [https://perma.cc/V4BW-MVCT]; Julie de Bailliencourt, *Meet TikTok's European Safety Advisory Council*, TIKTOK (Mar. 1, 2021), https://newsroom.tiktok.com/en-gb/tiktok-european-safety-advisory-council [https://perma.cc/MDQ9-NFZG].

66. *Introducing the Twitch Safety Advisory Council*, TWITCH BLOG (May 14, 2020), https://blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/ [https://perma.cc/2ZXF-BVJQ].

67. Nick Pickles, *Strengthening our Trust and Safety Council*, TWITTER BLOG (Dec. 13, 2019), https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html [https://perma.cc/R9B9-V3A3]; Patricia Cartes, *Announcing the Twitter Trust & Safety Council*, TWITTER BLOG (Feb. 9, 2016), https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html [https://perma.cc/L62R-AF54].

amici curiae fashion.[68] Since 2017, Facebook has had a formal team exclusively tasked with managing relationships with civil society organizations and academics to gather their input for policy development.[69]

These models have different features. TikTok and Twitch rely on very small advisory councils. According to what participants report in the interviews, YouTube engages civil society with an eye toward getting specific knowledge or areas of expertise, rather than using it as an opportunity to collect input from a broad group of actors as a standardized phase of their policy development process. At Facebook, consultation with civil society is a key step in the development of all changes to content moderation rules.[70] Dedicated staff consult with a broad group of people around the world who bring diverse perspectives.[71]

Unlike a notice-and-comment process, these participatory processes are not open to the public. Rather, input is solicited from specific stakeholders. This means that access is heavily controlled by the company. Also different from notice-and-comment rulemaking, Facebook does not promise to account for all the interests expressed by civil society. No specific guidance to weigh in stakeholders' "comments" exists. However, because it is a standardized phase of policy development and the outreach within the United States is broad and inclusive of diverse viewpoints, it is the closest model encouraging institutional participation of civil society in rulemaking processes. Accordingly, this Article has a stronger focus on Facebook than on other less developed stakeholder engagement efforts.

Part II offers a set of hypotheses on how civil society participation has influenced content moderation. It starts by identifying the winners and losers of this process and it follows with possible accounts of how power is shaped in these governance mechanisms.

---

68. *Oversight Board Charter*, art. 3, FACEBOOK (Sept. 2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf [https://perma.cc/JLJ8-S84S].

69. Matthias Kettemann & Wolfgang Schulz, *Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System*, LEIBNIZ INSTITUTE FOR MEDIA RESEARCH HANS-BREDOW-INSTITUT 23 (Jan. 2020), https://leibniz-hbi.de/uploads/media/default/cms/media/k0gjxdi_AP_WiP001InsideFacebook.pdf [https://perma.cc/JFV9-CVUB].

70. *Facebook Community Standards*, META, https://www.facebook.com/communitystandards/stakeholder_engagement [https://perma.cc/8S4V-KW3W] (last visited May 29, 2022) ("These standards are based on feedback from people and the advice experts in fields like technology, public safety[,] and human rights."); *How Stakeholder Engagement Helps Us Develop the Facebook Community Standards*, META (Jan 19, 2022), https://transparency.fb.com/policies/improving/stakeholders-help-us-develop-community-standards/ [https://perma.cc/DSD3-ZS8F].

71. *See generally* Kettemann & Schulz, *supra* note 69, at 17.

II. HOW IT HAS WORKED: THE EVOLUTION OF CONTENT MODERATION

A. *Outcomes*

In its short history, content moderation has evolved through various trends. Klonick describes how a few loose standards that could fit in one page at Facebook evolved into extensive granular rules,[72] although many of them are still vague.[73] Evelyn Douek identifies two additional patterns of change.[74] First, a proportional—instead of categorical—approach to content moderation has increasingly given more importance to values other than freedom of expression, such as safety and public health.[75] Second, companies more openly acknowledge that errors are unavoidable in content moderation at scale.[76] Therefore, instead of avoiding all false positives, which has been the preference of First Amendment doctrine,[77] companies are aiming at figuring out what error rates they prefer. Similarly, Zittrain describes a shift from a rights-based approach, where protecting freedom of expression was the paramount priority, to a public health approach, more concerned with balancing the benefits and risks of speech for society.[78]

Content moderation rules at all mainstream platforms have imposed increasing restrictions on freedom of expression.[79] Hate speech policies are a good place to look at these developments because they have been an important site for political struggle in the last decade.

It is well-known that the United States has afforded broad protections to extremist or hate speech.[80] In important ways, Facebook's rules have always widely diverged from First Amendment doctrine. Charlotte Willner, an early employee at Facebook, describes these initial standards: "If it

72. Klonick, *supra* note 1, at 1631.

73. *See, e.g.*, *SR Report 2018*, *supra* note 63, ¶ 26 ("Company prohibitions of threatening or promoting terrorism, supporting or presaging leaders of dangerous organizations . . . are . . . excessively vague." (footnotes omitted)).

74. Evelyn Douek, *Governing Online Speech: From "Posts-as-Trumps" to Proportionality and Probability*, 121 COLUM. L. REV., 759, 759 (2021).

75. *See id.* at 770.

76. *See id.* at 796–99.

77. *See, e.g.*, N.Y. Times Co. v. Sullivan, 376 U.S. 254 (1964).

78. *See* John Bowers & Jonathan Zittrain, *Answering Impossible Questions: Content Governance in an Age of Disinformation*, 1 HARV. KENNEDY SCH. MISINFORMATION. REV. 1, 5 (2020); Jonathan Zittrain, *Three Eras of Digital Governance* (Sept. 15, 2019) (unpublished manuscript), available at https://ssrn.com/abstract=3458435 [https://perma.cc/N8B6-8YKN].

79. Jillian C. York & Corynne McSherry, *Content Moderation is Broken. Let Us Count the Ways.*, ELEC. FRONTIER FOUND. (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways [https://perma.cc/Z4SX-4XM6] ("[T]hanks to growing pressure from governments and some segments of the public to restrict various types of speech, [content moderation] has also become more pervasive and aggressive[.]").

80. *See* Brandenburg v. Ohio, 395 U.S. 444, 449 (1969); Nat'l Socialist Party of Am. v. Village of Skokie, 432 U.S. 43, 43–44 (1977).

makes you feel bad in your gut, then go ahead and take it down."[81] These standards loosely prohibited "things like Hitler and naked people."[82] As the platform grew in size and became closer to resembling the public square, those standards did not converge with the First Amendment, but have evolved further apart.

Since May 2018, Facebook has changed its hate speech community guidelines twenty times.[83] All the changes have either made a rule more granular or restricted more content. The first kind of changes (more granular) are usually triggered by moderators identifying a lack of consistency in the application of a rule, or a trend in the application of the rule that does not respect the policy rationale.[84] The second kind of changes (stricter) are often triggered either by an outside report, an event covered in the media, or detection that hateful content that can be defined with a degree of precision is prevalent on the platform.[85]

Examples of the latter include the following. In August 2018, Facebook added cases as a protected category and new additional instances of "dehumanizing speech" were included in the list of forbidden content. In December 2019, more instances of dehumanizing speech and comparisons were prohibited. In June 2020, the firm deleted humor as an exception to the rule in the policy rationale.[86] In August 2020, combating harmful stereotypes was added to the policy rationale and two instances of forbidden dehumanizing comparisons were included: caricatures of Black people in the form of blackface and Jewish people running the world or major institutions. In October of the same year, Facebook banned "deny[ing] or distort[ing]" information about the Holocaust.[87]

81. Klonick, *supra* note 1, at 1631 (citing Telephone Interview with Dave Willner, Former Head of Content Policy, Facebook & Charlotte Willner, Former Safety Manager, User Operations, Facebook (Mar. 23, 2016)).

82. *Id.*

83. The company has made its rules and subsequent amendments public only since 2018. *See* Monika Bickert, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, META (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards [https://perma.cc/TLT6-B4M3]. Since then, the policy was changed in May 2018, August 2018, March 2019, July 2019 (twice), August 2019, October 2019, December 2019, February 2020, March 2020, June 2020, July 2020, August 2020, September 2020, October 2020, November 2020, January 2021, June 2021, October 2021, and November 2021. *See Hate Speech (Recent Updates), Community Standards*, META, https://transparency.fb.com/policies/community-standards/hate-speech/ [https://perma.cc/98MA-AJQG] (last visited July 14, 2022).

84. Interview with participant #57, Facebook employee (Sept. 28, 2020) [hereinafter Interview #57].

85. *Id.*

86. *See infra* Section III.A.1.

87. Monika Bickert, *Removing Holocaust Denial Content*, META (Oct. 12, 2020), https://about.fb.com/news/2020/10/removing-holocaust-denial-content/ [https://perma.cc/K 9QX-Z9ZN]; *see also* Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech but not Black Children*, PROPUBLICA (June 28, 2017, 5:00 AM), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.

Other changes have not been as significant, but rather have introduced more precise vocabulary. Examples include a clarification in the definition of exclusion in Tier 3 (Facebook organizes hate speech violations in three tiers in order of severity) from July 2019; refinement of the language in Tier 2 in July 2019;[88] and clarifications on how to differentiate generalizations (disallowed) and behavioral statements (allowed) in August 2019.[89] The only exception in this trend toward stricter rules is a change in March 2019. Facebook added an exception to the policies: "People sometimes express contempt in the context of a romantic breakup. Other times, they use gender-exclusive language to control membership in a health or positive support group, such as a breastfeeding group for women only. We allow the content but expect people to clearly indicate their intent." This change added new contexts in which prima facie violating content is now permissible.

Twitter has followed a similar trend. In July 2019, Twitter expanded the rules against hateful conduct to include language that dehumanizes others on the basis of religion or caste. In March 2020, the firm expanded the rule to include language that dehumanizes on the basis of age, disability, or disease. In December 2020, they added new protected categories: national origin, race, and ethnicity.[90] In January 2021, the company included a ban on posts that communicate hope an entire protected category dies as a result of a serious disease (previously only hoping that one individual person dies was banned). This amendment also clarified that inciting violence includes wishing, hoping, or calling for serious harm.[91] Similarly, YouTube's two major changes to their hate speech policies in 2020 expanded sanctions against repetitions of insults or abuses targeting groups based on protected categories.[92]

---

cc/2JSA-RVNN] (narrating how the discussions about Holocaust denial began and evolved inside Facebook).

88. *See Product Policy Forum Minutes*, FACEBOOK (Apr. 23, 2019), https://about.fb.com/wp-content/uploads/2018/11/ppf-final-deck_04.23.2019-1.pdf [https://perma.cc/237E-PNWN].

89. *See Product Policy Forum Minutes*, FACEBOOK (May 21, 2019), https://about.fb.com/wp-content/uploads/2018/11/ppf-final-deck_05.21.2019.pdf [https://perma.cc/KWX5-MH5U].

90. *See* Twitter Safety, *supra* note 64.

91. *See* La Letra Chica (@laletrachicatv), TWITTER (Jan. 19, 2021), https://coda.io/d/Letra-Chica-Twitter_dXmBknXgkmZ/Politica-relativa-a-las-conductas-de-incitacion-alodio_su_nC#_lu7I5 [https://perma.cc/K52H-2H5N]. *See generally* LETRA CHICA, https://letrachica.digital/ [https://perma.cc/2PE4-64SM] (last visited May 29, 2022) (tracking changes to YouTube, Facebook, and Twitter's policy automatically).

92. LETRA CHICA, https://letrachica.digital/ [https://perma.cc/2PE4-64SM] (last visited May 29, 2022).

B. *Winners and Losers*

I categorize civil society actors in two ideal types: speech "regulationists" and speech "protectionists." Speech regulationists maintain that the First Amendment—particularly in cases dealing with hateful, violent, and "silencing speech"—prioritizes the free speech of privileged speakers at the expense of the free speech or even the safety of disadvantaged or marginalized people.[93] On the contrary, speech protectionists generally believe that First Amendment doctrine ensures that the speech of marginalized speakers is not suppressed.

These are ideal types or stylized categories. I do not intend to overstate their differences: many organizations want platforms to allow more of some types of speech and less of other types. Fundamentally, almost everyone involved in these debates wants freedom of speech to be a means for self-expression and to fight injustices. Policy differences are often about the best methodology to achieve those ends. Still, these categories are helpful because they track the conflicts that surround changes to rules and how the policy debate has evolved. The categorization also responds to how the Facebook Content Policy Stakeholder Engagement team designs its outreach strategy. When discussing a new rule, the team looks for representatives of the different values that Facebook upholds: voice, on the one hand, and the counter-values of safety, dignity, authenticity, and privacy, on the other.[94] In practice, that often means identifying advocates for rules that are more protective of freedom of expression, and advocates more concerned with safety and dignity issues.[95]

To a large extent, the disagreements about rules has been a disagreement among liberal and leftist organizations. Politically conservative organizations have participated in content moderation debates, but for the most part have focused on legal reform and corporate enforcement of their own rules, rather than the content of the rules themselves. For instance, some Conservative figures have been concerned with platforms' alleged bias against them, despite the lack of evidence showing that this bias in fact exists.[96] Mainstream Conservative organizations such as the

93. For earlier formulations of these positions, see generally MARY MATSUDA, CHARLES L. LAWRENCE III, RICHARD DELGADO & KIMBERLÉ WILLIAMS CRENSHAW, WORDS THAT WOUND: CRITICAL RACE THEORY, ASSAULTIVE SPEECH, AND THE FIRST AMENDMENT (1993); Owen M. Fiss, *Free Speech and Social Structure*, 71 IOWA L. REV. 1405 (1986).

94. Monika Bickert, *Updating the Values that Inform Our Community Standards*, META (Sept. 12, 2019), https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/ [https://perma.cc/GW7L-DB7E].

95. Interview #57, *supra* note 84.

96. *See, e.g.*, Jon Kyl, *Covington Interim Report*, FACEBOOK NEWSROOM (Aug. 20, 2019), https://about.fb.com/wp-content/uploads/2019/08/covington-interim-report-1.pdf [https://perma.cc/K46J-B9HR]; Paul Barrett & J. Grant Sims, *False Accusation: The Unfounded Claim that Social Media Companies Censor Conservatives*, N.Y.U. STERN CTR. FOR BUS. & HUM. RTS. (Feb. 2021), https://static1.squarespace.com/static/5b6df958f8370af3217d4178/t/60187b5f45762e708708c8e9/1612217185240/NYU+False+Accusation_2.pdf [https://perma.cc/P5JV-4EKD].

Cato Institute and the Mercatus Center at George Mason University have resisted this narrative and have been staunch supporters of Section 230.[97] But Republican politicians have taken up this concern that they are being discriminated against at congressional hearings and have introduced proposals to amend Section 230 of the Communications Decency Act.[98] Likewise, The Heritage Foundation has echoed the concerns about bias against Conservatives.[99] But it has articulated its concerns mostly as a problem in the enforcement of rules, rather than in relation to any specific substantive standard.[100]

As is predictable from the content of the rules, regulationists often describe this evolution of rules as highly positive, while protectionists are largely dissatisfied. A regulationist advocate for victims of online sexual violence says that rules have changed very favorably: "The difference between now and then is night and day."[101] Members of the leading organizations of Change the Terms, a coalition that fights for more restrictions on hateful content, evaluate their campaigns as a success at the level of rules. A participant says, "I think what you will find over the last two years is that there is a lot of progress among some of the companies . . . . And you do see some changes that are more closely aligning with our Change the Terms model policies."[102] Her overall view of the engagements in the last few years is that "what you will see is not that they don't want to adopt the policies or that they're really pushing back but that the companies are progressing."[103]

For stakeholders pushing for regulation of harmful content, companies have been responsive in many cases: "If you look over the last five years and the policies that they have put in place, I would argue that we,

97. John Samples, *Armslist and Bias against Conservatives Online*, CATO BLOG (May 14, 2019, 1:44 PM), https://www.cato.org/blog/armslist-bias-against-conservatives-online [https://perma.cc/H3VS-QHYR]; Brent Skorup & Jennifer Huddleston, *The Erosion of Publisher Liability in American Law, Section 230, and the Future of Online Curation*, MERCATUS CTR. GEORGE MASON UNIV. (July 11, 2019), https://www.mercatus.org/publications/regulation/erosion-publisher-liability-american-law-section-230-and-future-online [https://perma.cc/Y9AT-M32N].

98. *See, e.g.*, *Senator Hawley Introduces Legislation to Amend Section 230 Immunity for Big Tech Companies*, JOSH HAWLEY U.S. SENATOR FOR MISSOURI (June 19, 2019), https://www.hawley.senate.gov/senator-hawley-introduces-legislation-amend-section-230-immunity-big-tech-companies [https://perma.cc/X4UM-SZGX].

99. *See, e.g.*, Kay C. James, Commentary, *Social Media's Preemptive Spiking of New York Post Story Shows Bias Against Conservatives Continues*, HERITAGE FOUND. (Oct. 22, 2020), https://www.heritage.org/civil-society/commentary/social-medias-preemptive-spiking-new-york-post-story-shows-bias-against [https://perma.cc/7W6A-3ST8].

100. *See id.* ("Key takeaways . . . . Social media platforms need to leave their biases out of their business models and apply their rules and standards fairly and impartially.").

101. Interview with participant #31 (Oct. 19, 2020) [hereinafter Interview #31].

102. Interview with participant #27 (Aug. 10, 2020) [hereinafter Interview #27].

103. *Id.*

and other groups, and our direct negotiations with them, led to that policymaking quicker than any threat of oversight from the government, at least in the [United States]."[104]

Despite agreement on the progress at the level of rules, participants express a deep frustration with these platforms. Speech regulationists have been successful in terms of how rules have evolved. However, it is unclear to them how companies operate in the background. Referring to an independent civil rights audit that Facebook concluded in 2020 that showed the inconsistent enforcement of hate speech policies, a participant evaluates that "the audit told us what we already knew: that rules support white supremacists . . . . Banning is only good as enforcement."[105] The mismatch between public announcements and media reports increases that distrust. Because little information is available about how rules are enforced at the system level, activists do not have the tools or the normative framework to assess the recurrent instances of violating speech found on platforms.

On the protectionist side, a member of a leading tech policy organization compares the first years when "these companies were little babies" to the present time.[106] Ten to fifteen years ago "[social media companies' employees] were very like-minded kind of people" who "really wanted to know what we thought."[107] Now, the organization can get companies to correct enforcement mistakes (i.e., when the company has taken down content that does not violate their own policies or copyright law), but on the policy level "even if we're good friends with their lead counsel or whatever, we just don't have the same ability to influence them the way that we used to."[108] Despite increasing convergence in rules across platforms, a difference between companies still exists. In her view, "there's a lot of people at YouTube and there's a lot of people at Twitter still that really are trying to promote free speech. I think Facebook is kind of gone. They're gone. We don't have a lot of influence with them."[109]

Staunch free speech advocates have expressed frustration over corporate restrictions on speech,[110] opposition to asking companies to police

104. Interview with participant #21 (Aug. 8, 2020) [hereinafter Interview #21].

105. Interview with participant #20 (July 17, 2020) [hereinafter Interview #20].

106. Interview with participant #22 (July 29, 2020) [hereinafter Interview #22].

107. *Id.*

108. *Id.*

109. *Id.*

110. *See, e.g.*, Daphne Keller, *Facebook Restricts Speech by Popular Demand*, ATLANTIC (Sept. 22, 2019), https://www.theatlantic.com/ideas/archive/2019/09/facebook-restricts-free-speech-popular-demand/598462/ [https://perma.cc/YS43-SSBQ].

speech,[111] and concern about over-enforcement of corporate rules.[112] A participant who heads the online speech work of an organization based in Washington, D.C. says their organization does not "want Nazis on platforms either," but "hate speech is very hard to define even for governments," so "platforms should look at the specific problem they want to solve" and avoid current blanket bans that cover unclear categories of speech.[113]

C. *Possible Explanations*

Many factors can explain this trend toward greater restrictions on online speech. It could be a response to the influence of other stakeholders; it could be explained through public choice theory; it could respond to a general shift in social preferences driven by technological developments.

First, this section asks whether civil society matters at all or whether the changes to content moderation rules respond solely, or primarily, to business reasons or direct government actions. Next, this section examines two existing hypotheses for how voices calling for stricter rules have become louder in recent years. Both accounts highlight an important aspect of the evolution of content moderation. However, they are incomplete and overlook an unexplored dimension: the role of law in shaping the dynamics and outcomes of civil society participation.

1. *Does Civil Society Participation Matter?*

Many factors could explain the evolution of content moderation rules independently of civil society. As platforms penetrated global markets, companies have encountered different normative requirements and preferences. European institutions have played a leading role in pushing toward restrictions on content that is legal in the United States.[114] Scholars

111. *See, e.g.*, Corynne McSherry, *Corporate Speech Police Are Not the Answer to Online Hate*, ELECTRONIC FRONTIER FOUND. (Oct. 25, 2018), https://www.eff.org/deeplinks/2018/10/corporate-speech-police-are-not-answer-online-hate [https://perma.cc/PMY8-T5P9].

112. *See, e.g*, Paresh Dave, *YouTube Reversal Bans Holocaust Hoaxers, Stops Pay for Borderline Creators*, REUTERS (June 5, 2019 12:51 PM), https://www.reuters.com/article/us-alphabet-youtube-hatespeech/youtube-reversal-bans-holocaust-hoaxers-stops-pay-for-borderline-creators-idUSKCN1T623X [https://perma.cc/U9SD-S9KB] (quoting Jennifer Granick, from the American Civil Liberties Union, expressing concerns that YouTube will make mistakes and over-censor content when applying a ban on Holocaust denial).

113. Interview with participant #01 (Apr. 1, 2020) [hereinafter Interview #01].

114. *See, e.g.*, *Digital Service Act*, *supra* note 7 (requiring social media companies to block or remove content that violates the German Criminal Code's restrictions on defamatory and hate speech); Case C-18/18, Glawischnig-Piesczek v. Facebook Ir. Ltd., 2019 E.C.R., ¶ 53 (Oct. 3, 2019) (allowing state members to order the global takedown of content deemed illegal according to domestic law).

have described in depth how governments around the world have leveraged companies to delete content.[115]

In the United States, despite First Amendment limitations to speech regulation, government officials have not been shy to make express demands that companies take down legal speech.[116] Advertisers have pushed companies in the same direction.[117] With many actors pushing toward stricter rules, it is possible that these normative outcomes respond exclusively to factors other than the advocacy of U.S. civil society.

One key question then is whether civil society participation matters at all.[118] At first glance, the response is "sometimes." One participant in this research compares conversations with tech firms to being in *Groundhog Day*, the movie in which Bill Murray finds himself reliving the same day over and over again.[119] Others report concrete cases in which their suggestions were directly reflected in amendments to rules, in decisions about specific instances of content, or in the enforcement guidelines and training materials designed for content moderators.

The limited literature addressing this question reflects the mixed experiences the participants report. Matthias Kettemann and Wolfgang Schulz studied the development of content rules at Facebook.[120] They

115. *See* James Boyle, *A Nondelegation Doctrine for the Digital Age?*, 50 DUKE L.J. 5, 6 (2000); Michael D. Birnhack & Niva Elkin-Koren, *The Invisible Handshake: The Reemergence of the State in the Digital Environment*, 8 VA. J.L. & TECH. 1, 6 (2003); Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105, 106 (2016); Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 96 (2006); Balkin, *supra* note 5; Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296 (2014); Rabea Eghbariah & Amre Metwally, *Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service*, 23 YALE J.L. & TECH 543 (2021).

116. *See* Rebecca Heilweil, *Amy Klobuchar Takes Aim at 12 Vaccine Misinformation Influencers*, VOX (Apr. 19, 2021, 2:26 PM), https://www.vox.com/recode/2021/4/19/22388231/amy-klobuchar-vaccine-misinformation-facebook-twitter-superspreaders [https://perma.cc/2H58-GWUF] (describing a letter senators have written to Facebook and Twitter urging them to take action against specific accounts that distribute falsehoods about vaccines); *Warren, Carper, Whitehouse, Schatz Call on Mark Zuckerberg and Facebook to Stop the Deliberate Spread of Climate Disinformation on the Company's Social Media Platforms*, ELIZABETH WARREN (July 16, 2020), https://www.warren.senate.gov/oversight/letters/warren-carper-whitehouse-schatz-call-on-mark-zuckerberg-and-facebook-to-stop-the-deliberate-spread-of-climate-disinformation-on-the-companys-social-media-platforms [https://perma.cc/5JA5-6KTC].

117. *See, e.g.*, *WFA and platforms make major progress to address harmful content*, WORLD FED'N OF ADVERTISERS (Sept. 23, 2020), https://wfanet.org/knowledge/item/2020/09/23/WFA-and-platforms-make-major-progress-to-address-harmful-content [https://perma.cc/4FLS-U4ZH].

118. *See* Lucian Bebchuk & Roberto Tallarita, *The Illusory Promise of Stakeholder Governance*, 106 CORNELL L. REV. 91, 110 (2020) (arguing that civil society lacks the power to override shareholders' interests).

119. Interview with participant #02 (Apr. 8, 2020) [hereinafter Interview #02]; GROUNDHOG DAY (Columbia Pictures 1993).

120. Kettemann & Schulz, *supra* note 69.

observed that "external stakeholders are extensively referred to in the discourse on policy changes" and that the engagements between Facebook and external stakeholders "were characterized by a sincere interest in exchanging views on policy development, mutual appreciation, and an open exchange of ideas that contributed positively to the policy development process."[121] At the same time, David Kaye highlighted in his 2018 report to the Human Rights Council that "[p]articipants in consultations consistently raised concerns that companies failed to engage adequately with users and civil society, particularly in the global South."[122]

Tracing whether it is stakeholders' input that drives a policy change is not always possible. Participants complain about the lack of feedback after the consultations and many times they are unaware of whether their suggestions were adopted or not.[123] In some instances, participants have the impression that staff who engage with them have no internal power[124] or that input is solicited even though the policy change has already been decided.[125] However, participants can also point to specific examples where their input was reflected in the policy outcome.

A member of the TikTok U.S. Content Advisory Council, commenting on the new policy on deep fakes rolled out in 2020,[126] stated that "when [TikTok was] trying to write their deep fake rules, we really made quite significant changes to that policy. My recollection is they adopted just about everything we recommended."[127] An organization that advised Twitter about their policy on violent threats reported:

> That was a really positive example of having an impact . . . . They came to us. We told them all the things that were wrong with it. They went back and revised it. Then they came back to

121. *Id.* at 23.

122. *SR Report 2018*, *supra* note 63, ¶ 54.

123. *See* Interview with participant #28 (Sept. 24, 2020) [hereinafter Interview #28] (describing the follow-up emails that Facebook usually sends out after a consultation and expressing concerns that the emails are often insufficiently precise to track how her own input was processed).

124. *See* Interview with participant #41 (June 12, 2020) [hereinafter Interview #41].

125. Interview #31, *supra* note 101 ("Twitter tends to tell the Trust and Safety Council 'Hey, we rolled out this thing. What do you think? And by the way it doesn't matter what you think because we already pulled it out.'"). *See also* Louise Matsakis, *Twitter Trust and Safety Advisers Say They're Being Ignored*, WIRED (Aug. 23, 2019, 2:50 PM), https://www.wired.com/story/twitter-trust-and-safety-council-letter/ [https://perma.cc/WF2Y-NMW8].

126. Vanessa Pappas, *Combating Misinformation and Election Interference on TikTok*, TIKTOK NEWSROOM (Aug. 5, 2020), https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok [https://perma.cc/5897-NMXB].

127. Interview with participant #29 (Sept. 23, 2020) [hereinafter Interview #29].

> address our concerns, which is really kind of an ideal model of engagement.[128]

Another participant discusses a Twitter initiative to call for submissions to comment on a proposed new policy on dehumanizing content.[129] She thinks "the proposal was awfully broad."[130] Feedback from the public did end up persuading Twitter that the policy should be narrower.[131]

In August 2020, Facebook added a ban on content depicting blackface. According to the minutes of the meeting discussing the policy, Facebook consulted with sixty stakeholders including academics and civil society organizations around the world.[132] A member of the Facebook Content Policy Stakeholder Engagement team says that the proposed ban was strongly supported by a solid majority of stakeholders.[133] Indeed, in the minutes of the meeting, one of the listed "pros" of the new policy on one the slides is that the change is "grounded in stakeholder feedback."[134] In 2021, the Facebook Oversight Board confirmed this when reviewing the blackface ban in the case about Zwarte Piet (Black Pete).[135] Ultimately, the Board allowed Facebook to adopt the ban even though the rule, according to the Board itself, was not compatible with international human rights law.

These experiences follow findings in the literature about notice-and-comment showing that participation does influence substantive outcomes.[136] The question, then, is not whether participation of civil society matters, but when and why it does. Why do companies care and consider what civil society has to say instead of dismissing it? How does that feedback help companies maximize profits?

Part of the literature and public conversation assumes that companies' interests are simple and straightforward: they promote whatever con-

---

128. Interview with participant #16 (June 11, 2020) [hereinafter Interview #16].

129. *See* Twitter Safety, *supra* note 64 (describing new policy).

130. Interview with participant #46 (June 18, 2020).

131. *See* Kate Conger, *Twitter Backs Off Broad Limits on 'Dehumanzing' Speech*, N.Y. TIMES (July 9, 2019), https://www.nytimes.com/2019/07/09/technology/twitter-ban-speech-dehumanizing .html [https://perma.cc/9YEY-ZL3A].

132. *Product Policy Forum Minutes*, FACEBOOK (Aug. 11, 2020), https://about.fb.com/wp-content/uploads/2018/11/PPF_08.11.2020_Harmful-Stereotypes.pdf [https://perma.cc/MG3R-TL FE].

133. Interview with participant #60 (Sept. 28, 2020) [hereinafter Interview #60].

134. *Product Policy Forum Minutes*, *supra* note 132.

135. *Case decision 2021-002-FB-UA*, OVERSIGHT BD. (Apr. 13, 2021), https://oversightboard.com/decision/FB-S6NRTDAJ/ [https://perma.cc/23F2-QPS6].

136. *See* Susan Webb Yackee, *Sweet-Talking the Fourth Branch: The Influence of Interest Group Comments on Federal Agency Rulemaking*, 16 J. PUB. ADMIN. RSCH. & THEORY 103, 105 (2006); Amy McKay & Susan Webb Yackee, *Interest Group Competition on Federal Agency Rules*, 35 AM. POL. RSCH. 336 (2017).

tent retains the attention of their users.[137] The actual picture, however, is more complex. Companies pursue competing and conflicting interests: promoting engagement and maximizing screen time,[138] ensuring that users return to use their services,[139] positioning themselves as socially responsible businesses,[140] accounting for advertisers' interests,[141] preventing media outrage,[142] responding to government pressure,[143] and operating within their technical possibilities.[144] All these interests coexist and push rules in different directions. In that context, the rules that maximize revenue are unstable, indeterminate, and ever changing.

The way Facebook employees describe the company's content moderation goals reflects the indeterminacy of social media companies' rules preferences. One participant says that Facebook's primary value is freedom of expression but also (in the same interview) that the company would "obviously like to eliminate all hateful content" if operational barriers did not exist.[145] What this indeterminacy shows is that there is no one single rule that the business model dictates.

The launch of the Facebook Oversight Board confirms that the business model is to a large extent indifferent to any specific rule. This body makes policy recommendations to Facebook.[146] These recommendations

137. *See generally* ADAM ATLER, IRRESISTIBLE: THE RISE OF ADDICTIVE TECHNOLOGY AND THE BUSINESS OF KEEPING US HOOKED (2017); TIM WU, THE ATTENTION MERCHANTS: THE EPIC SCRAMBLE TO GET INSIDE OUR HEADS (2016) (discussing addictive design and incentives to maximize the amount of time users spend on their devices).

138. *See, e.g.*, Wu, *supra* note 137.

139. *See, e.g.*, Nick Clegg, *You and the Algorithm: It Takes Two to Tango*, MEDIUM (Mar. 31, 2021), https://nickclegg.medium.com/you-and-the-algorithm-it-takes-two-to-tango-7722b19aa1c2 [https://perma.cc/H4DB-BB9J] ("The company's long-term growth will be best served if people continue to use its products for years to come. If it prioritized keeping you online an extra 10 or 20 minutes, but in doing so made you less likely to return in the future, it would be self-defeating.").

140. *See, e.g.*, Klonick, *supra* note 1, at 1627; Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 53 (2018).

141. *See, e.g.*, Susan Wojcicki, *Letter from Susan: Our 2021 Priorities*, YOUTUBE OFF. BLOG (Jan. 26, 2021), https://blog.youtube/inside-youtube/letter-from-susan-our-2021-priorities/ [https://perma.cc/RYX5-S8LA] ("Our policies are designed to protect our YouTube community against abuse and bad actors and also to make sure we are able to keep advertisers coming back to YouTube . . . .").

142. *See, e.g.*, Bickert, *supra* note 10, at 265; Mike Ananny & Tarleton Gillespie, *Public Platforms: Beyond the Cycle of Shocks and Exceptions*, OXFORD INTERNET INST. (2017), http://blogs.oii.ox.ac.uk/ipp-conference/sites/ipp/files/documents/anannyGillespie-publicPlatforms-oii-submittedSept8.pdf [https://perma.cc/8PPD-KXKT].

143. *See, e.g.*, Boyle, *supra* note 115.

144. *See, e.g.,* Natasha Duarte, Emma Llanso & Anna Loup, *Mixed Messages? The Limits of Automated Social Media Content Analysis*, 81 PROC. MACHINE LEARNING RSCH. 106 (2018) (describing some of the technical limitations to moderate content).

145. Interview #57, *supra* note 84.

146. *Oversight Board Bylaws*, Art. 2, META (Jan. 2020), https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf [https://perma.cc/T9C9-522H].

are not binding, but Facebook has fully or partially implemented many of the suggestions of the Board.[147] The company's decision to renounce some of its regulatory power in favor of opening itself to different policy options shows both that it can grow under multiple different sets of rules and that the perceived legitimacy of its rules may be even more important than the rules themselves.

Even if retaining users' attention is the core of the business model, content moderation rules are not the only tool to achieve that. Rather, appeasing public outrage may be a low-cost concession for avoiding threats over privacy regulation and antitrust enforcement that would more directly put the business model under pressure.[148] Indeed, scholars agree that companies have changed their content moderation practices in response to public opinion.[149] Engagement with civil society is a tool to manage the risk of public criticism.[150]

Understanding how civil society participation advances corporate interests helps explain why it has an effect on policy development, as well as its most important limitations. As a risk management tool, it is likely that it is more impactful on the aspects of content moderation that are visible and, therefore, closer to public scrutiny. In fact, civil society participation has had an impact on content moderation rules, which are public and subject to criticism. It is possible that the same impact does not translate to more opaque areas of content moderation such as enforcement and algorithmic distribution of content because they are less transparent, harder to assess, and their opacity may lower the risk of public criticism. It is also likely that companies are more responsive to the positions of those advocates who are more effective risk makers. I will revisit this point in Part III where I discuss how advocates' engagement with legal discourse has contributed to their ability to position themselves as risk makers.

2. *Public Choice Theory*

Civil society actors have more often urged companies to delete more lawful-but-awful content—not less.[151] Noah Feldman posits that Mançur Olson's theory of collective action can explain this asymmetry:

147. *Oversight Board Recommendations*, FACEBOOK TRANSPARENCY CTR., https://transparency.fb.com/oversight/oversight-board-recommendations/ [https://perma.cc/CUG9-8YW2] (last visited July 14, 2022).

148. *See* JULIE COHEN, BETWEEN TRUTH AND POWER: THE LEGAL CONSTRUCTIONS OF INFORMATIONAL CAPITALISM 51 (2019) (deconstructing the legal systems that enable platforms' business models).

149. *See, e.g.*, Keller, *supra* note 110; Ananny & Gillespie, *supra* note 142; Noah Feldman, *A Supreme Court for Facebook* (Jan. 1, 2018), *in* GLOBAL FEEDBACK AND INPUT ON THE FACEBOOK OVERSIGHT BOARD FOR CONTENT DECISIONS app.101 (June 27, 2019); York & McSherry, *supra* note 79.

150. The position of Director of Stakeholder Engagement at Facebook grew from a previous position as Risk Manager when the workload was enough to create a whole team. *See* Interview #60, *supra* note 134.

151. *See* Keller, *supra* note 6, at 13.

> There's nothing inherently wrong with urging private companies to regulate offensive or potentially dangerous speech. The problem is the asymmetry between those who seek censorship and those who favor freedom. Those calling for restricting speech tend to be highly motivated, focused on their issue of choice, whether it be racist speech or sexist speech or just speech they really believe should not exist. Yet those who favor free speech typically hold that interest in a more diffuse way. They believe in freedom of expression for everybody, but are not specifically concentrated on any particular utterance or viewpoint.
>
> In a lobbying battle between a concentrated group and a diffuse interest, the concentrated advocacy group almost always wins . . . . So when advocates pressure Facebook or Google or Twitter to ban certain speech, they have a good chance of getting through to the companies. No one is pushing hard on the other side of the door.[152]

Olson's theory theorizes that large groups with diffuse interests may be systematically overpowered by smaller groups with concentrated and well-defined interests.[153] Those who directly suffer or acutely benefit from a policy will have stronger incentives to lobby for or against it than a constituency of individuals who are less directly impacted.

In the regulation of harmful speech, this could manifest in those being directly harmed having stronger incentives to organize and lobby than those who favor the protection of that speech. The reason is that speech protectionists frequently will not be protecting their own speech but the speech of others that they often dislike. For example, the concentrated and well-defined interest to have companies delete content such as white supremacy or the video depicting a terrorist attack in Christchurch, New Zealand would facilitate coordination among the interested actors. On the other hand, for those advocates who have an interest in protecting this type of speech, the social costs may be more diffuse and therefore less likely to transform into action.

This theory explains some aspects of the dynamics among civil society groups. On the one hand, the leader of a racial justice organization narrates the beginnings of their advocacy around content moderation. They

152. Feldman, *supra* note 149, at app. 101.

153. *See generally* Mançur Olson, THE LOGIC OF COLLECTIVE ACTION: PUBLIC GOODS AND THE THEORY OF GROUPS (1971). Olson does not have a definitive account of how his theory applies to non-economic interests. On the one hand, he recognizes that the framework might be inadequate to explain philanthropic or "noneconomic" lobbies. *Id.* at 160. On the other hand, Olson believes the theory can "cover all types of lobbies." *Id.* at 159. *See also* Kapczynski, *supra* note 21, at 811 n.6 (discussing the scope of interests in public choice theory); Yochai Benkler, *Through the Looking Glass: Alice and the Constitutional Foundations of the Public Domain*, L. & CONTEMP. PROBS. 173, 196 (2003) (thinking through Olson's framework about economic, social, political, and moral costs).

first turned to Facebook to complain about the white supremacist comments users would leave on the organization's Facebook page.[154] Another organization was initially concerned with racist comments that the organization's staff and other community members were receiving on social media, especially on Facebook.[155] In these examples, victims or their close representatives become advocates for rules that would directly address the harm they experience online.

In contrast, an advocate with a feminist organization complains about how other feminists have spent time on getting gendered-based slurs banned on Facebook. She says, "I am sorry but calling someone names is OK."[156] However, organizing a campaign to protect gendered slurs on platforms is not their top priority.[157] Public choice helps make sense of this asymmetry: those directly harmed by harmful content have a more acute interest in avoiding that content than those who would rather see some of that content protected.

However, public choice theory does not explain why groups have mobilized and organized differently before government institutions than before corporate actors. In 1964, the American Civil Liberties Union (ACLU) successfully defended Clarence Brandenburg, a leader in the Ku Klux Klan. The Supreme Court agreed that the First Amendment protects speech no matter how racist or supportive of violence when it does not directly encourage people to immediately engage in unlawful action.[158] When a neo-Nazi group wanted to march through the Chicago suburb of Skokie, it was the ACLU that represented the interests of the National Socialist Party of America. These cases highlight the organization's long-standing commitment to defend freedom for "speech we hate."[159] More recently, free speech organizations have been vocal and effective in fighting for freedom of online expression before courts and governmental bodies. To name one example, the ACLU has successfully litigated over the constitutionality of statues banning the nonconsensual distribution of intimate images that the organization considers overbroad.[160]

---

154. Interview #20, *supra* note 105.

155. *See* Interview with participant #13 (May 27, 2020) [hereinafter Interview #13].

156. Interview with participant #07 (May 6, 2020) [hereinafter Interview #07].

157. *Id.*

158. Brandenburg v. Ohio, 395 U.S. 444, 449 (1969).

159. *See generally* Philippa Strum, WHEN THE NAZIS CAME TO SKOKIE: FREEDOM FOR SPEECH WE HATE (1999).

160. *See, e.g.*, Complaint at 5, Antigone Books v. Horne, No. 2:14-cv-02100-PHX-SRB (D. Ariz. 2015), https://www.aclu.org/legal-document/Antigone-books-v-horne-complaint [https://perma.cc/VZ36-Y582]; Lee Rowland, *VICTORY! Federal Judge Deep-Sixes Arizona's Ridiculously Overbroad 'Nude Photo' Law*, ACLU (July 10, 2015, 6:45 PM), aclu.org/blog/free-speech/internet-speech/victory-federal-judge-deep-sixes-arizonas-ridiculously-overbroad [https://perma.cc/D6LK-BLKW]; *Free Speech and Media Groups Applaud Governor's Veto of Overbroad "Revenge Porn" Bill*, ACLU (June 21, 2016), https://www.aclu.org/press-releases/free-speech-and-me

Before private regulators, free speech advocates have not achieved—nor fought as vigorously for—equivalent outcomes. Public choice theory is unable to explain this difference.

3. *A Technology-Driven Shift in Social Norms*

It could be that public pressure to delete more content does not stem from better organized lobbying from one side, as public choice theory would posit, but from a broad, new consensus in normative preferences driven by technological innovation. Douek describes a "cultural reevaluation" of how online speech should be moderated.[161] Douek exemplifies the point with a New York Times opinion piece that declares the "idea that platforms like Twitter, Facebook and Instagram should remove hate speech is relatively uncontroversial."[162] As Daphne Keller notes, "Even for committed free-expression advocates, it is not clear that requiring platforms to preserve all legal speech is in the public interest."[163]

A few factors explain how technology could have driven this "change of heart."[164] First, most scholars and users agree that some types of content, though legal, diminish the value of platforms for expressive activities. The clearest example is the regulation of spam. Spam is not necessarily illegal content, but something that "nearly all users agree should go."[165] Likewise, many consider sexual content as similar to spam because its presence on mainstream platforms risks disturbing the user experience.[166] Famously, the original motivation for Congress to pass the Communications Decency Act was to incentivize platforms to restrict pornography.[167] The

dia-groups-applaud-governors-veto-overbroad-revenge-porn-bill [https://perma.cc/A6V8-S54X].

161. Douek, *supra* note 74, at 778–79.

162. *Id.* at 780; Brittan Heller, *Is This Frog a Hate Symbol or Not?*, N.Y. TIMES (Dec. 24, 2019), https://www.nytimes.com/2019/12/24/opinion/pepe-frog-hate-speech.html [https://perma.cc/C72G-WUEB].

163. Keller, *supra* note 6, at 13.

164. Quinta Jurecic, *Gab Vanishes, and the Internet Shrugs*, LAWFARE (Oct. 29, 2018, 6:37 PM), https://www.lawfareblog.com/gab-vanishes-and-internet-shrugs [https://perma.cc/7HTV-7SQH].

165. TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA, 217 n.22 (2018); *see also* FINN BRUNTON, SPAM: A SHADOW HISTORY OF THE INTERNET (2013).

166. *See, e.g.*, Molly Wood, *Here's the Real Deal with Section 230*, MARKETPLACE (Dec. 7, 2020), https://www.marketplace.org/shows/marketplace-tech/section-230-protections-online-publishers-facebook-youtube-trump-defense-spending-bill-veto-censorship-accusations/ [https://perma.cc/4BE2-8276] ("[I]f you've been following the development of Parler, the so-called free-speech app . . . you'll have seen that the app is being flooded with porn and escort services. That was not only inevitable but also shows why totally unmoderated platforms often end up being kind of bad for business."). However, platforms that allow nudity and sexual content, such as Twitter, show that sexual content does not necessarily lower the value of platforms or worsen the user experience.

167. *See* Klonick, *supra* 1, at 1605; KOSSEFF, *supra* note 25, at 10.

assumption is not that sexual content ought to be banned outright, but that it is appropriate to silo it to designated digital zones. Therefore, the proper functionality of platforms requires that companies have the capacity to exclude some forms of legal content.[168]

Second, scholars theorize that online speech causes greater harm and, thus, requires stricter regulations than offline speech.[169] Gender-based harassment is a prime example that illustrates this point. One of the basic premises in Danielle Citron's seminal work about online violence against women is that the nature of the new medium created greater harms that require proportional responses.[170] The accumulation of offensive comments that can reach users from multiple unconnected and distant people transforms some comments that would be permissible as isolated offline instances into unbearable online experiences.[171]

More recently, the anxieties around the role of social media in the Christchurch terrorist attack,[172] the genocide in Myanmar,[173] and in the election of Donald Trump in 2016 may have had a long-lasting impact on liberals' commitment to freedom of expression.[174] It is plausible that social media's role as a scapegoat for Donald Trump's election and his use of platforms during his presidency have made strict speech regulation attractive as a handy fix to deeper societal issues.[175]

These factors can explain an emerging shift in social consensus. However, even if almost everyone agrees that platforms should be able to restrict speech somewhat more aggressively than the state has been allowed to, there is no consensus about how broad that gap ought to be. As previ-

168. *See* GILLESPIE, *supra* note 165.

169. *See infra* Section III.A.

170. *See* Danielle Keats Citron, *Cyber Civil Rights*, 89 B.U. L. REV. 61, 63 (2010) (arguing that regulation must be responsive to the fact that internet's affordances make various types of content particularly harmful as users can aggregate their efforts to harm others and disaggregate their offline presence from their online identities). *See also* Soraya Chemaly, *Demographics, Design, and Free Speech: How Demographics Have Produced Social Media Optimized for Abuse and the Silencing of Marginalized Voices*, *in* FREE SPEECH IN THE DIGITAL AGE 150 (Susan Brison & Katharine Gelber eds., 2019).

171. *See* Viktorya Vilk, Elodie Vialle & Matt Bailey, *No Excuse for Abuse: What Social Media Companies Can Do Now to Combat Online Harassment and Empower Users*, PEN AMERICA, https://pen.org/report/no-excuse-for-abuse/ [https://perma.cc/2ADB-CRC4] (last visited July 14, 2022).

172. *See, e.g.*, Kevin Roose, *A Mass Murder of, and for, the Internet*, N.Y. TIMES (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html [https://perma.cc/7Z4A-WS8Z].

173. *See, e.g.*, Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. TIMES (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/yanmar-facebook.html [https://perma.cc/75SU-JTAJ].

174. *See, e.g.*, Nathaniel Persily, *Can Democracy Survive the Internet?*, 28 J. DEMOCRACY 63 (2017).

175. *See* BENKLER, FARIS & ROBERTS, *supra* note 5 (challenging the idea that social media platforms played a main role in producing misinformation during the 2016 presidential campaign).

ously discussed, speech protectionist advocates are unsatisfied and concerned with the evolution of content moderation rules. The next Part argues that their relative failure can be better understood as a failure to imagine a normative lobbying agenda to participate meaningfully in the governance of online speech.

III. THE ASYMMETRIC U.S. CIVIL SOCIETY

This Part offers an alternative hypothesis to explain the asymmetry between advocacy groups that seek restrictions on harmful speech and those who favor its protection. This account explores the role of law in shaping advocates' lobbying agendas and their ability to position themselves as risk makers.

A. *Speech Regulationists*

Speech regulationists sometimes attribute their normative preferences to technological realities. For example, because online speech has more reach, it is more dangerous and has to be strictly regulated; in this new context, the American preferred remedy for bad speech—more speech—does not work because platforms are optimized for engagement, not truth-seeking.[176] Leading speech regulationists, however, often believe that First Amendment doctrines are also inadequate to regulate offline speech. The head of a racial justice organization, for instance, explains that though the First Amendment is sanctimonious for other civil society organizations, it has not served people of color in the United States well.[177]

The speech regulations that these advocates support would be highly unlikely to survive First Amendment scrutiny before the Supreme Court if they were adopted by the state. But speech regulationists have found in social media corporations a more receptive regulator. In a highly concentrated media environment, activists can ask a handful of companies to adopt the rules that the government refuses to pass. This strategy has been an escape from the First Amendment laissez faire doctrines.

This is also why regulationists have resisted the analogies between social media and state actors that could lead to the expansion of the scope of application of the First Amendment to these spaces.[178] For those who be-

176. *See, e.g.*, Mary Anne Franks, *The Free Speech Black Hole: Can the Internet Escape the Gravitational Pull of the First Amendment?*, KNIGHT FIRST AMEND. INST. (Aug. 21, 2019), https://knightcolumbia.org/content/the-free-speech-black-hole-can-the-internet-escape-the-gravitational-pull-of-the-first-amendment [https://perma.cc/V7L2-3JPF] ("More and more online platforms are confronting the reality that the best answer to bad speech is not, in fact, more speech.").

177. Interview #13, *supra* note 155.

178. Interview #31, *supra* note 101 ("Let's say there's a hate speech policy that it [sic] could either be broader or narrower. And you'll have people on the Council who will say, look, if we're going to treat this like a First Amendment free speech kind of issue, we should be as narrow as possible . . . . And then you have

lieve that the First Amendment laissez faire doctrines are inadequate to promote an ideal speech environment, expanding those standards to the social media sphere is not promising. Thus, they have instead embraced the state action requirement of the First Amendment, which means that social media platforms are not bound by First Amendment doctrines.[179]

The First Amendment state action requirement defines the contours of a space that seems lawless and ought to be ruled instead by social corporate responsibility. Racial and social justice groups have effectively organized to demand that gatekeepers exercise their power responsibly. A notable example is the "Change the Terms" coalition, which brings together organizations such as the Center for American Progress, Color of Change, Free Press, Lawyers' Committee for Civil Rights under Law, MediaJustice, Muslim Advocates, National Hispanic Media Coalition, and the Southern Poverty Law Center.[180] In 2018, they launched a set of model policies for internet companies to reduce hateful activities.[181] The document presenting the model policies explains that "[b]ecause internet tools are largely owned and managed by the private sector and not government, these corporations must be part of the solution to address the promulgation of hateful activities online."[182] The philosophy "can implies ought" underlies the demands. The assumption is that what is "ought" is to prevent the harms caused by the combination of user-generated speech and platform architecture.

However, embracing the state action requirement has not meant that speech protectionists have abstained from legal discourse. On the contrary, this strict separation between government and corporate actors has offered an opportunity to articulate their demands in the terms of more fruitful legal frameworks: civil rights and conceptions of the right to freedom of expression that have failed before United States courts.

1. *Civil Rights*

In 2010, Citron proposed cyber civil rights as a promising legal framework to respond to the harmful consequences of speech on the internet. Its appeal to speech regulationists resides in the fact that it moves the con-

---

others, and I count myself in this group, who would say, 'Look, you are not a government entity and you are shaping your own community, the way that you want to shape it. There's no reason for you to follow along any kind of legalistic sense of what has to be left alone. You can decide whether or not you want this kind of content on your platform.'").

179. *See* Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019) ("[M]erely hosting speech by others is not a traditional, exclusive public function and does not alone transform private entities into state actors subject to First Amendment constraints.").

180. *About*, CHANGE THE TERMS, https://www.changetheterms.org/about/ [https://perma.cc/5QNL-JDT3] (last visited July 14, 2022).

181. *See generally Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities*, *supra* note 58.

182. *Id.* at 2.

versation away from the permissibility of speech under First Amendment doctrine and puts a stronger priority on the protection of victims of harassment and discrimination.[183] Another attractive feature is that U.S. scholars are more used to applying civil rights horizontally among private parties.

Conducted at the behest of civil rights organizations such as MediaJustice, Color of Change, and the Anti-Defamation League, Facebook audits between 2018 and 2020 illustrate how the civil rights framework functions. The audit concluded with a report that includes several recommendations for content moderation. In the introduction to the final report, the auditors expressed concern about "[t]he prioritization of free expression over all other values, such as equality and non-discrimination . . . ."[184] Accordingly, recommendations were directed at broadening the prohibitions against discriminatory content.

For example, the auditors recommended that Facebook expand its policy against white nationalism to prohibit content which expressly praises, supports, or represents white nationalist or separatist ideology even if it does not explicitly use those terms.[185] They recommended that Facebook remove humor as an exception to its prohibition on hate speech "because humor was not well-defined and was largely left to the eye of the beholder."[186] Facebook followed through by replacing humor with a narrower definition of satire.[187]

On this point, a decision by the Facebook Oversight Board discussing satirical content highlights how the civil and human rights frameworks are mobilized to emphasize different normative commitments. The case originated with Facebook's decision to delete an image depicting Turkey having to choose between only two options: "The Armenian Genocide is a lie" and "The Armenians were terrorists who deserved it." Although a clear satire of Turkey's denialist position toward the Armenian genocide, Facebook justified the decision in a blog post explaining that "[w]e do not allow hate speech on Facebook, even in the context of satire, because it

183. *See* Citron, *supra* note 170, at 66.

184. *See* LAURA W. MURPHY & MEGAN CACACE, FACEBOOK'S CIVIL RIGHTS AUDIT—FINAL REPORT 9 (July 8, 2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/95TU-TKF6].

185. *See id.* at 50.

186. *Id.* at 44.

187. *See id.* Satire is defined "as content that 'includes the use of irony, exaggeration, mockery and/or absurdity with the intent to expose or critique people, behaviors, or opinions, particularly in the context of political, religious, or social issues. Its purpose is to draw attention to and voice criticism about wider societal issues or trends.'" *Id.* The exception is not included in the public language of Facebook's Community Standards. *See Case decision 2021-005-FB-UA*, OVERSIGHT BD. (May 20, 2021), https://www.oversightboard.com/decision/FBRZL57QHJ [https://perma.cc/ELE9-JTSR] [hereinafter *Case decision 2021-005-FB-UA*] ("While [the humor] exception was removed, [Facebook] kept a narrower exception for satire that is currently not communicated to users in its Hate Speech Community Standard.").

creates an environment of intimidation and exclusion, and in some cases, may promote real-world violence."[188] In its submission to the Oversight Board, Facebook stated that its rule against hate speech had previously included an exception for humor, but that the exception was removed in response to the recommendations of the civil rights audit report in 2020.[189] However, the Oversight Board noted that in the civil rights audit report the company had disclosed that it had kept a narrower exception for satirical content.

Ultimately, a majority of the Oversight Board found that the content fell under two exceptions to Facebook's hate speech policies: content that includes hate speech to condemn it, and satire. In its recommendations to Facebook and under an international human rights law assessment, the Oversight Board recommended that Facebook "make sure it has adequate procedures in place to assess satirical content."[190]

I do not mean to imply that in this specific case the same or the opposite conclusions could not have been reached using civil rights or international human rights law language. Indeed, a minority of the Oversight Board reached the opinion that Facebook's original decision to delete the content was correct under a human rights law framework. Likewise, from a position mostly concerned with preventing discrimination on the platform, protecting this post condemning Turkey's position on the Armenian genocide could be of the utmost importance.

Rather, the two frameworks pursue different aims in the broader picture. The two sides of the debate reflect different preferences on which side to err. At the scale of content that Facebook moderates, gray cases are numerous, and moderators have little time to make decisions. In this context, the civil rights audit recommended eliminating the exception for humorous content to avoid that moderators apply it "far too frequently."[191] That is, in gray cases, the exception could lead moderators to allow speech that the auditors would consider impermissible. Facebook claimed to defend the elimination of the exception on that basis: "creating a definition for what is perceived to be funny was not operational for Facebook's at-scale enforcement."[192] Note that the auditors' recommendation was not to draw a narrower exception for satirical content, but to remove it.[193]

On the other hand, the Facebook Oversight Board recommended building mechanisms to ensure that satirical content was not deleted.

188. *Oversight Board Selects Case on a Comment Related to the Armenian People and the Armenian Genocide*, META (Mar. 2, 2021), https://about.fb.com/news/2021/03/oversight-board-selects-case-on-comment-related-to-the-armenian-people-and-the-armenian-genocide/ [https://perma.cc/2RQF-BPMM].

189. *Case decision 2021-005-FB-UA*, *supra* note 187.

190. *Id.*

191. *See* MURPHY & CACACE, *supra* note 184, at 44.

192. *Case decision 2021-005-FB-UA*, *supra* note 187.

193. *See* MURPHY & CACACE, *supra* note 184, at 44.

That is, the Board's preference was that Facebook ensure that in gray cases, the firm will make sure not to delete speech that could be justifiable as satirical. The human rights framework could go further and recommend that Facebook overturn a policy that prohibits speech that does not incite immediate illegal action altogether.[194] But the issue was not raised in this case.

In any event, the controversy around exceptions to the ban on hate speech illustrates that advocates mostly concerned with restricting discriminatory speech have resorted to civil rights as the appropriate framework for regulation. Those who have been mostly concerned with avoiding falsely positive cases of hate speech have relied on other legal frameworks.[195]

2. *Silencing Speech*

When discussing the compatibility between civil rights and freedom of expression, Citron argues that civil rights enhance, rather than undermine, free speech.[196] This understanding of freedom of expression can be traced back to Owen Fiss's influential work on freedom of expression and social groups.[197] Even before the internet era, Fiss argued that regulating "silencing" speech could maximize the overall amount of expression or the number of viewpoints represented in the public sphere. For example, in the context of campaign finance, an overwhelming amount of speech by one candidate can drown down the speech of its competitors. Discriminatory, hateful, and offensive speech can lead other speakers to abandon the conversation.[198] Silencing speech inhibits speakers from subordinated groups to participate in public discourse by undermining their status as equal members of society in such a way that they are either deterred from speaking or their speech is taken less seriously by others.[199]

The Supreme Court expressly rejected this rationale to regulate speech in the campaign finance case *Buckley v. Valeo*,[200] when saying that "the concept that government may restrict the speech of some elements of

194. *See infra* Section III.B.3 (discussing this position).

195. *See infra* Part III (describing these alternative frameworks).

196. *See* Danielle Keats Citron, *Restricting Speech to Protect It*, *in* FREE SPEECH IN THE DIGITAL AGE 122 (Susan J. Brison & Katharine Gelber eds., 2019).

197. OWEN FISS, THE IRONY OF FREE SPEECH 13 (1996); Fiss, *supra* note 93, at 1407. The question of speech and disadvantaged groups could also be framed as a tension between free speech and equality. This is how the issue was framed in the work of Catharine MacKinnon. CATHARINE A. MACKINNON, ONLY WORDS, 69 (1993). However, contemporary speech regulationists have adopted the framing proposed by Owen Fiss as a tension between freedom of speech and freedom of speech. *See, e.g.*, Citron, *supra* note 170, at 98 (citing Owen Fiss's work on silencing speech).

198. *See generally* Danielle Keats Citron & Jonathon W. Penney, *When Law Frees Us to Speak*, 87 FORDHAM L. REV. 2317 (2019).

199. FISS, *supra* note 197, at 16.

200. 424 U.S. 1 (1976).

our society in order to enhance the relative voice of others is wholly foreign to the First Amendment."[201] But in the internet age, the theory has gained new support. A report by PEN America on online abuse targeted at journalists explains that "[o]nline abuse is intended to intimidate and censor. When voices are silenced and expression is chilled, public discourse suffers."[202] Relying on the same framework, in *Cyber Civil Rights*, Citron argued that banning some types of hardcore sexual content was good for everyone because as such materials become increasingly available, "increasing numbers of parents will restrict or deny their children's Web access and other adults will turn away from the Internet in disgust."[203]

Social media companies have adopted this theory as an almost self-evident, valid rationale to moderate content and design platforms' architecture. Twitter's Vijaya Gadde said in 2015 that "[f]reedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up."[204] Facebook's recently updated rationale for its hate speech policy states that "[w]e believe that people use their voice and connect more freely when they don't feel attacked on the basis of who they are. That is why we don't allow hate speech on Facebook."[205] Companies have made slow progress in rolling out features that help users protect their privacy and avoid "silencing speech." For example, in 2013, Twitter integrated a reporting button into their apps.[206] In 2014, it included a mute feature to hide undesired comments.[207]

---

201. *Id.* at 48–49.

202. Vilk, Vialle & Bailey, *supra* note 171.

203. Citron, *supra* note 170, at 85. To illustrate how advocates adopt a civil right or a human rights perspective to justify different normative commitments, note the difference between this passage and the Facebook Oversight Board's repeated statement that international human right law protects offensive speech. *See Case Decision 2021-005-FB-UA*, *supra* note 187; *Case decision 2021-002-FB-UA*, OVERSIGHT BD. (Apr. 13, 2021), https://oversightboard.com/decision/FB-S6NR TDAJ/ [https://perma.cc/ZZ7A-WRPD] ("The UN Human Rights Committee has made clear the protection of Article 19 extends to expression that may be considered 'deeply offensive.'").

204. Vijaya Gadde, Editorial, *Twitter Executive: Here's How We're Trying to Stop Abuse While Preserving Free Speech*, WASH. POST (Apr. 16, 2015), https://www.washingtonpost.com/posteverything/wp/2015/04/16/twitter-executive-heres-how-were-trying-to-stop-abuse-while-preserving-free-speech/ [https://perma.cc/TVQ8-CEMQ].

205. *Hate Speech (Recent Updates)*, META: FACEBOOK CMTY. STANDARDS (Apr. 2022), https://www.facebook.com/communitystandards/recentupdates/hate_speech [https://perma.cc/8H55-D2SX].

206. *See* Alexander Abad-Santos, *Twitter's 'Report Abuse' Button Is a Good, But Small, First Step*, ATLANTIC (July 31, 2013), theatlantic.com/technology/archive/2013/07/why-twitters-report-abuse-button-good-tiny-first-step/312689/ [https://perma.cc/YZ6G-23JY].

207. *See* Paul Rosania, *Another Way to Edit Your Twitter Experience: With Mute*, TWITTER BLOG (May 12, 2014), blog.twitter.com/official/en_us/a/2014/another-way-to-edit-your-twitter-experience-with-mute.html [https://perma.cc/M4W6-CNZH]. *But see* Vilk, Vialle & Bailey, *supra* note 171 (listing many design features

Besides enhancing the agency of users to decide what speech to encounter, other calls have been more specifically targeted at community standards to prohibit "silencing" content. In 2013, Women, Action and the Media (WAM!), The Everyday Sexism Project, and activist Soraya Chemaly issued a letter to Facebook demanding the company address gender-based hate speech.[208] As a result, Facebook committed to reviewing their hate speech standards.[209] Years later, in November 2018, Facebook banned targeting individuals with terms related to sexual activity. And in February 2020, they expanded the ban to female-gendered cursing terms.[210]

This rationale to regulate speech is based on the effects of the speech and does not draw clear lines among different types of speech that silences others. Therefore, companies have used it to justify bans on content that inhibits the participation of other users without any analysis of whether those "chilling" effects are enough to restrict speech. For example, Facebook justifies its broad ban on nudity with the justification that the content can make (some) women feel uncomfortable and drive them away from participating on the platform.[211] Likewise, the policy rationale for Facebook's rule on sexual solicitation explains that the ban aims at preventing trafficking and nonconsensual sexual acts, as well as addressing the fact that "some audiences within our global community may be sensitive to this type of content, and it may impede the ability for people to connect with their friends and the broader community."[212] The use that companies have made of the concept of "silencing speech" to publicly justify their own business decisions show the success that the rationale has gained in the platform era.

Overall, advocates and scholars have been successful in introducing Fiss's theory of silencing speech into the mainstream discourse about on-

---

and privacy settings that companies should adopt to enhance the experience of journalists online).

208. *See* Rachel Walden, *Campaign Against Gender-Based Hate Speech on Facebook—Activists Win!*, OUR BODIES OUR SELVES BLOG (May 29, 2013), https://www.ourbodiesourselves.org/2013/05/campaign-against-gender-based-hate-speech-on-facebook-activists-win/ [https://perma.cc/XY79-EXHD].

209. Marne Levine, *Controversial, Harmful and Hateful Speech on Facebook*, FACEBOOK (May 28, 2013), https://www.facebook.com/notes/facebook-safety/controversial-harmful-and-hateful-speech-on-facebook/574430655911054 [https://perma.cc/WM9A-GQB3].

210. *See Bullying and Harassment (Recent Updates)*, META: FACEBOOK CMTY. STANDARDS (Feb. 2020), https://www.facebook.com/communitystandards/recentupdates/bullying/ [https://perma.cc/6H3T-3M9H].

211. *Adult Nudity and Sexual Activity*, META: FACEBOOK CMTY. STANDARDS https://www.facebook.com/communitystandards/adult_nudity_sexual_activity [https://perma.cc/W3S5-ZBFF] (last visited May 29, 2022) ("We restrict the display of nudity or sexual activity because some people in our community may be sensitive to this type of content.").

212. *Sexual Solicitation*, META: FACEBOOK CMTY. STANDARDS, https://www.facebook.com/communitystandards/sexual_solicitation [https://perma.cc/ZQQ7-23BS] (last visited May 29, 2022).

line speech governance. However, it is worth noting that companies have adopted the theory with one essential modification. Companies usually regulate speech targeted at protected categories symmetrically. That is, they ban "silencing" speech attacking both dominant and subordinated groups.[213] Fiss's theory and other foundational theories in favor of hate speech regulation do not support—in fact, are at odds with—that symmetrical approach.[214]

3. *"Free Speech Is Not the Same as Free Reach"*[215]

A common position emerged in recent years that is best captured by Renee DiResta's words, "free *speech* does not mean free *reach*."[216] This phrase aims at distinguishing users' speech from companies' decisions about how speech circulates. Indeed, users choose what to say online, but platforms determine "which speech is amplified and which is suppressed."[217] As previously discussed, the harmful effects of online speech come only partially from what users say.[218] Platforms' architecture influences how prevalent speech is and what audiences it reaches. Algorithmic recommendations on YouTube and the ranking of content on Facebook and Twitter newsfeeds may push people to increasingly extremist, radicalizing content, echo chambers, or connections with attackers across the world.[219] It is intuitive that the harm speech can create depends both on the size of the audience and the accumulation of offending speech. Accordingly, the distinction between speech and reach tries to capture that the problem is not speech itself but platforms' architecture.[220]

---

213. *See, e.g.*, Simon Van Zuylen-Wood, *"Men Are Scum": Inside Facebook's War on Hate Speech*, VANITY FAIR (Feb. 26, 2019), https://www.vanityfair.com/news/2019/02/men-are-scum-inside-facebook-war-on-hate-speech [https://perma.cc/5WZ7-VBGP].

214. FISS, *supra* note 197, at 21; MATSUDA, LAWRENCE, DELGADO & CRENSHAW, *supra* note 93, at 36.

215. Renee DiResta, *Free Speech Is Not the Same as Free Reach*, WIRED (Aug. 30, 2018, 4:00 PM), https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/ [https://perma.cc/SL5H-F7UA].

216. *Id.*

217. Jameel Jaffer, *Facebook and Free Speech Are Different Things*, KNIGHT FIRST AMEND. INST. (Oct. 24, 2019), https://knightcolumbia.org/content/facebook-and-free-speech-are-different-things [https://perma.cc/R2WL-SXHA].

218. *See supra* Section II.C.3.

219. *See Facebook: Turn Off the Algorithms*, ACCOUNTABLE TECH, https://accountabletech.org/campaign/facebook-turn-off-the-algorithms/ [https://perma.cc/F97W-7E QK]; Jonas Kaiser, Adrian Rauchfleisch & Yasodara Córdova, *Fighting Zika With Honey: An Analysis of YouTube's Video Recommendations on Brazilian YouTube*, 15 INT'L J. COMMC'N 1244 (2021); CASS SUNSTEIN, #REPUBLIC: DIVIDED DEMOCRACY IN THE AGE OF SOCIAL MEDIA (2017).

220. *See* Jaffer, *supra* note 217 ("But the more powerful critique of Facebook . . . is that the prevalence of false, hateful, and sensational speech on the platform is a function not of the inherent value or persuasiveness of users' speech but of Facebook's decisions.").

As Keller shows, this normative distinction does not necessarily reflect how the First Amendment treats speech and its distribution when it comes to state regulation.[221] Although the mere ability to speak is different from the distribution of speech, laws restricting the distribution of speech and laws banning speech are subject to the same strict scrutiny. To this point, the Supreme Court said in *United States v. Playboy*,[222] "[i]t is of no moment that the statute does not impose a complete prohibition. The distinction between laws burdening and laws banning speech is but a matter of degree. The Government's content-based burdens must satisfy the same rigorous scrutiny as its content-based bans."[223]

Here, once again, lobbying companies enabled a reconceptualization of the right to freedom of expression. The strategy has been successful to the extent that it has permeated corporate discourse and become somehow common-sense in conversations about content governance.[224] Companies have embraced this way of thinking about freedom of expression. Most notably, in his 2018 blueprint, Mark Zuckerberg presented as a positive—and not often criticized—approach that the company would limit the distribution of "borderline" content. Borderline content by definition does not violate Facebook's community standards, but it is somehow close to it. One of the examples Zuckerberg provided was photos "with revealing clothing or sexually suggestive positions." This content does not violate Facebook's ban on nudity, but the company curbed its distribution.[225]

Often, content distribution is discussed as a merely technical question. If the problem is not what users say but how companies distribute it, perhaps a technical discussion about platforms' architecture can replace the normative one. The unspoken assumption tends to be that responsible companies will curb the distribution of problematic speech, regardless of its legal status or other normative considerations.

---

221. Daphne Keller, *Amplification and Its Discontents: Why Regulating the Reach of Online Content Is Hard*, KNIGHT FIRST AMEND. INST. (June 8, 2021), https://knightcolumbia.org/content/amplification-and-its-discontents [https://perma.cc/5UBH-K7S4].

222. 529 U.S. 803 (2000).

223. *Id.* at 812.

224. *See, e.g.*, Evelyn Douek, *What Facebook Did for Chauvin's Trial Should Happen All the Time*, ATLANTIC (Apr. 21, 2021), https://www.theatlantic.com/ideas/archive/2021/04/facebook-should-dial-down-toxicity-much-more-often/618653/ [https://perma.cc/3A8B-KATU] (arguing that if social media companies can reduce the prevalence of hate speech on their platforms, they should adopt that approach); Julia Alexander, *YouTube Claims Its Crackdown on Borderline Content is Actually Working*, VERGE (Dec. 3, 2019, 12:00 PM), https://www.theverge.com/2019/12/3/20992018/youtube-borderline-content-recommendation-algorithm-news-authoritative-sources [https://perma.cc/8Q9S-AUEP].

225. *See* Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK, https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/2XAH-C36R] (last modified May 5, 2021).

Overall, the projects aimed at reducing distribution of problematic speech would face severe challenges if articulated as legal reform projects in the United States.[226] However, advocates have successfully framed it as the responsible normative approach that platforms should follow. Regulationists have adopted this approach not only publicly but also in their conversations with companies.[227]

B. *Speech Protectionists*

Like the speech regulationists' efforts, speech protectionists' projects have been deeply grounded on the First Amendment's state action requirement. Their ideal has been a state that stays out of speech regulation, while the market provides multiple platforms with competing speech rules. The private side of this ideal would produce information laboratories.[228] The model derives from Justice Louis Brandeis's concept of democracy laboratories[229] and quite literally recreates Justice Oliver Wendell Holmes's marketplace of ideas.[230] From this perspective, freedom of expression thrives when self-regulated markets offer distinct systems that compete for truth production.[231] This vision of free speech as a competitive market is fundamentally at odds with a preference for any specific substantive rule. It gives almost total deference to the owners of each space to choose the rules that make up the service offered to the public.

Accordingly, one participant explains that when a platform reaches out to consult on a new policy, there are very few areas in which the organization has a pre-determined policy they would recommend.[232] Instead, their goal is to understand the needs of the platform, the environment and social norms the company wants to foster, and *then* advise on policies that will help the platform fix its problems.

As some platforms control too-large portions of information flows, the ideal stumbles. With the rise of powerful gatekeepers, speech protectionists have started seeing these rules as too important to be left to the market

226. *See* Keller, *supra* note 221 (providing a detailed account of such challenges and offering regulatory options that emphasize content-neutral goals such as privacy and competition).

227. *See, e.g.*, Interview #27, *supra* note 102 (saying that when corporate staff justify not deleting discriminatory content because it does not violate their policies, their response is that "free speech is not free reach" and its distribution should be limited).

228. *See* COHEN, *supra* note 148, at 76 (describing the information laboratories ideal).

229. New State Ice Co. v. Liebmann, 285 U.S. 262, 311 (1932) (Brandeis, J., dissenting).

230. Abrams v. United States, 250 U.S. 616, 630 (1919) (Holmes, J., dissenting).

231. *See* COHEN, *supra* note 148, at 75–107 (describing the "information laboratory" ideal).

232. Interview #01, *supra* note 113 (saying that the exception is policies that require users to use their real identities because "they do not work").

alone. Accordingly, advocates have pushed against increasingly restrictive rules.[233] For example, the same day that the Change the Terms coalition issued their model policy recommendations urging companies to address hate speech more aggressively, the Electronic Frontier Foundation (EFF) published a note against harnessing platforms' power to control speech.[234] The note highlighted their agreement with the demands for transparency and accountability, but pointed out that asking companies to be the speech police is "a profoundly dangerous idea."[235]

However, protectionists' position on private regulation has been ambivalent, or at least more ambivalent than their position on government regulation and more ambivalent than the position of speech regulationists. Even in its opposition to the Change the Terms demands, the EFF emphasized the dangers of over-enforcement and government intervention rather than relying on the classic defenses of hate speech that have defeated government censorship in courts.

As a specific example of the asymmetric approaches adopted by protectionists and regulationists, it is useful to compare the reaction to YouTube's decision to ban Holocaust denial of Jonathan Greenblatt, chief executive to the Anti-Defamation League, with the reaction of Jennifer Granick, a member of the ACLU. While the former advised Facebook to follow YouTube's example and adopt the same ban, the latter cautioned that "YouTube will make mistakes and over-censor," but "as a private company is well within its rights."[236] This example captures the core of the asymmetry: while Greenblatt has a specific rule that he would urge all services to adopt, the ACLU position focuses on potential errors in enforcement and does not urge the adoption of any specific balance of values.

This ambivalence is reinforced by the broad consensus that if companies allowed all legal speech, their services would become unusable.[237] That belief makes it necessary to acknowledge that companies cannot host all lawful speech. A line then must be drawn, but so far no protectionist project has found a principled way of establishing such a line in clear terms. That tension between concerns about how much lawful speech gatekeepers control and an ideal in which each space adopts its own rules

233. *See, e.g.*, Keller, *supra* note 110; Corynne McSherry, *Platform Censorship: Lessons from the Copyright Wars*, ELEC. FRONTIER FOUND. (Sep. 26, 2018), https://www.eff.org/deeplinks/2018/09/platform-censorship-lessons-copyright-wars [https://perma.cc/6778-VRKT]; Abdul Rahman Al Jaloud, Hadi Al Khatib, Jeff Deutch, Dia Kayyali & Jillian C. York, *Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content* (May 2019), https://www.eff.org/files/2019/05/30/caught_in_the_net_whitepaper_2019.pdf. [https://perma.cc/8HMW-TTZ3].

234. *See* McSherry, *supra* note 111.

235. *Id.*

236. Dave, *supra* note 112.

237. *See* GILLESPIE, *supra* note 165, at 5; Keller, *supra* note 6, at 13; Douek, *supra* note 74, at 768; Balkin, *supra* note 5, at 2026.

has left these advocates without strong tools to push social media companies in any direction.

I identify three projects that protectionists have tried to advance to counter the control over speech that giant platforms exercise. Here, I call them the Modern Public Square project, the Administrative Law project, and the Human Rights project.

1. *The Modern Public Square*

In *Packingham v. North Carolina*,[238] the Supreme Court referred to social media as "the modern public square."[239] In that spirit, scholars and litigants have explored the idea of extending users' First Amendment rights to online platforms. Ronald Krotoszynski synthesizes the underlying justification: "If a private company comes to possess monopoly power over an *essential* (rather than merely 'traditional') forum for democratic discourse, the First Amendment should impose limits on how and when the forum's owner may exercise the power to deny access to the forum for speech activity."[240]

The project builds upon a line of Supreme Court precedents on public forums and access duties starting in 1946 with *Marsh v. Alabama*[241] and ending in 1980 with *PruneYard Shopping Center v. Robins*.[242] In *Marsh*, the Supreme Court held that Grace Marsh, a member of the Jehovah's Witnesses, had a First Amendment right to distribute religious literature in the defendant's company-owned town. Justice Black argued that the accidental public or private ownership of the town did not impact the public's interest in the functioning of the community in such manner that the channels of communication remain free.[243] In a later case, *Amalgamated Food Employees Union Local 590 v. Logan Valley Plaza*,[244] the Court extended the reach of *Marsh*. Members of a labor union tried to access the Logan Valley shopping mall in order to picket a grocery store staffed with nonunion employees. *Logan Valley* was a different case from *Marsh* because the owners of the mall were not performing the "full spectrum of municipal powers."[245] Still, the *Logan Valley* Court ruled that the union members

238. 137 S. Ct. 1730 (2017).

239. *Id.* at 1737.

240. RONALD J. KROTOSZYNSKI, JR., THE DISAPPEARING FIRST AMENDMENT 51 (2019). Before social media, scholars envisioned analogous projects redesigning the public forum doctrine to make it applicable to the internet more broadly and other services. *See* David J. Goldstone, *The Public Forum Doctrine in the Age of the Information Superhighway (Where Are the Public Forums on the Information Superhighway?)*, 46 HASTINGS L.J. 335 (1995); Edward J. Naughton, *Is Cyberspace a Public Forum? Computer Bulletin Boards, Free Speech, and State Action*, 81 GEO. L.J. 409 (1992).

241. 326 U.S. 501 (1946).

242. 447 U.S. 74 (1980).

243. *See Marsh*, 326 U.S. at 507.

244. 391 U.S. 308 (1968).

245. Lloyd Corp., Ltd. v. Tanner, 407 U.S. 551, 569 (1972) (drawing this distinction).

had a right to access the property to exercise their free speech rights. The lack of an equally efficacious forum was a key factor in the case.[246]

The Supreme Court later limited *Logan Valley*[247] and ultimately overruled it.[248] In *Hudgens v. National Labor Relations Board*,[249] a union sought access to a shopping mall to picket a store located inside. This time, the Court sided with the owners of the shopping mall. Justice Potter Stewart stated that the First Amendment is "a guarantee only against abridgment by government, federal or state."[250] But *Hudgens* left a door open: Justice Stewart claimed that *in the absence of a statute* that limits the owner's rights to exclude others from its property, the Constitution itself does not "provide redress against a private corporation or person who seeks to abridge the free expression of others."[251] In *PruneYard*,[252] the Court walked through that open door. The Court upheld a California Supreme Court decision that read the California state constitution to require a private shopping mall to permit onsite protest activities.[253] Thus, although the First Amendment does not create a right to hold speech activities on private property open to the public, a statute that establishes that owners of such property must permit speech activities could be constitutional.[254]

On this basis, Krotoszynski has argued that Congress could impose must-carry duties on social media platforms.[255] More recently, Genevieve Lakier's work on the history of must-carry duties at different times of highly concentrated media environments in the United States shows that this type of regulation is not at all foreign to the American free speech tradition.[256] Eugene Volokh has recently examined the constitutionality and benefits of a common-carrier model for regulating the most powerful

246. *See Logan*, 391 U.S. at 324.

247. *See id.* (holding that a right to access private property is limited to circumstances where a nexus exists between a protest and the specific private property).

248. *See* Hudgens v. Nat'l Lab. Rels. Bd., 424 U.S. 507 (1976).

249. 424 U.S. 507 (1976).

250. *Id.* at 513.

251. *Id.*

252. 447 U.S. 74 (1980).

253. *Id.* at 88.

254. *See* KROTOSZYNSKI, *supra* note 240, at 62; Louis Michael Seidman, *The Dale Problem: Property and Speech Under the Regulatory State*, 75 U. CHI. L. REV. 1541, 1564 (2008).

255. *See* KROTOSZYNSKI, *supra* note 240, at 69–73. *See also* Michael C. Dorf, *Could Clarence Thomas Be Right About Twitter?*, JUSTIA: VERDICT (Apr. 14, 2021), https://verdict.justia.com/2021/04/14/could-clarence-thomas-be-right-about-twitter [https://perma.cc/742C-28T7].

256. *See* Lakier, *supra* note 9, at 2368–69. *See also* Genevieve Lakier & Nelson Tebbe, *After the "Great Deplatforming": Reconsidering the Shape of the First Amendment*, L. & POL. ECON. PROJECT (Mar. 1, 2021), https://lpeproject.org/blog/after-the-great-deplatforming-reconsidering-the-shape-of-the-first-amendment/ [https://perma.cc/W5QV-RVFA] (showing that precedent exists for regulating speech in spaces owned by private companies).

social media platforms.[257] Volokh argues that this model could possibly be appropriate for platforms' obligations to host content, while they could still compete for offering a desirable service to their users through content curation and algorithmic recommendation.[258]

In the absence of legislation, litigants have attempted to revive the reasoning of *Logan Valley* and have sought an easement right to speak on social media. All these attempts, to date, have failed.[259] In *Johnson v. Twitter*,[260] the California Superior Court did not consider Twitter a private shopping mall.[261] In *Prager v. Google*,[262] the Ninth Circuit did not regard YouTube as a "company town" or a state actor fulfilling a traditional function of the state.[263] In *Freedom Watch v. Google*,[264] the D.C. Circuit was clear that "the First Amendment 'prohibits only *governmental* abridgment of speech'" and "a private entity who provides a forum for speech is not transformed by that fact alone into a state actor."[265] Most recently, the District Court for the Northern District of California dismissed the claims of the Children's Health Defense, which alleged that Facebook had violated their free speech rights by limiting their ability to control their Facebook page and adding a label that fact-checks their claims about vaccines while also linking users to official vaccine information.[266] The court reasoned that Facebook had not acted together with any governmental agency and that the immunity provided by Section 230 does not provide sufficient encouragement to convert Facebook's private acts into state action.[267]

---

257. Eugene Volokh, *Treating Social Media Platforms Like Common Carriers?*, 1 J. FREE SPEECH L. 377, 382 (2021).

258. *Id.* at 410–11.

259. Keller, *supra* note 6, at 2. *See* Prager Univ. v. Google LLC, 951 F.3d 991, 999 (9th Cir. 2020) (rejecting Prager University's claim that YouTube is a public forum and finding that the state action doctrine bars constitutional review of the platform's content censoring); Buza v. Yahoo!, Inc., No. C 11–4422 RS, 2011 WL 5041174 (N.D. Cal. Oct. 24, 2011) (refusing to hold Yahoo!, a private company, liable for the suppression of free speech); Langdon v. Google, Inc., 474 F. Supp. 2d 622, 632 (D. Del. 2007) (dismissing the plaintiff's First Amendment claim because a private search engine is not a state actor subject to free speech guarantees).

260. No. 18CECG00078, LEXIS 8199, (Cal. Super. Ct. June 6, 2018).

261. *Id.* at *11.

262. 951 F.3d 991(9th Cir. 2020).

263. *Prager*, 951 F.3d at 998.

264. 816 F. App'x 497 (D.C. Cir. 2020).

265. *Id.* at 499 (quoting Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1928, 1930 (2019)).

266. Child. Health Def. v. Facebook, Inc., 546 F. Supp. 3d 909, 945 (N.D. Cal. 2021).

267. *Id.* at 930–31, 933. This argument about encouragement has also failed in other contexts as well. *See also* Divino Group LLC v. Google LLC, No. 19-cv-04749-VKD, 2021 WL, at *7 (N.D. Cal. Jan. 6, 2021); Newman v. Google LLC, No. 20-CV-04011-LHK, 2021 WL 2633423, at *10 (N.D. Cal. June 25, 2021); Eric Goldman, *Facebook Defeats Lawsuit By Publishers of Vaccine (Mis?)information–Children's Health Defense v. Facebook*, TECH. & MKTG. L. BLOG, (July

The First Amendment's state action requirement explains why these claims have failed before courts, but not why the project has not evolved into a lobbying agenda. As discussed above, the substantive projects of speech regulationists would very much likely fail before U.S. courts as well, but regulationists have still been able to engage corporations and the general public in their normative projects. In contrast, speech protectionists have not publicly argued or privately lobbied companies to follow First Amendment doctrines. The difference may be that for speech protectionists the barrier is not just legal but political.

In fact, the virtues of extending the First Amendment's protections of freedom of expression to the realm of social media have not gained broad consensus among scholars or speech protectionist advocates.[268] The position of leading speech protectionists is best reflected in the amicus brief the EFF submitted in *Prager v. Google* arguing that YouTube has a First Amendment right to moderate content as it sees fit.[269]

So far, the public–private divide has left speech protectionists with only two imaginaries available: the strict constitutional protections that the First Amendment provides against governmental interventions, and the competition among private actors for rules that serve their users best.[270] Without the possibility of transferring the public ideal into the private space, protectionists have not solidified a substantive agenda to lobby corporations. Without a vision for substantive rules, the most prominent way in which speech protectionists have advocated for some duties from gatekeepers has been to focus on procedural safeguards, to which I turn next.

2. *The Administrative Law Project*

The administrative law project comprises a series of proposals to incorporate administrative law principles to address the democratic deficits

2, 2021), https://blog.ericgoldman.org/archives/2021/07/facebook-defeats-lawsuit-by-publishers-of-vaccine-misinformation-childrens-health-defense-v-facebook.htm [https://perma.cc/H7X3-5632].

268. Balkin, *supra* note 5, at 2025 ("[T]he best alternative to this autocracy is not the imposition of First Amendment doctrines by analogy to the public forum or the company town."); Eric Goldman & Jess Miers, *Online Account Terminations/Content Removals and the Benefits of Internet Services Enforcing Their House Rules*, 1 J. FREE SPEECH L. 191, 196 (2021) (noting the adverse side-effects of must-carry rules and arguing that they would be counterproductive); Peters, *supra* note 3, at 2026 ("Imposing the same First Amendment doctrines that apply to municipalities to social media companies would quickly make these spaces far less valuable to end users, if not wholly ungovernable."); Christopher S. Yoo, *Free Speech and the Myth of the Internet as an Unintermediated Experience*, 78 GEO. WASH. L. REV. 697, 700–01 (2010) (arguing that editorial discretion bolsters desirable free speech values).

269. Brief of Amicus Curiae Electronic Frontier Foundation in Support of Defendants–Appellees at 2, Prager Univ. v. Google LLC, 951 F.3d. 991 (2020) (No. 18-15712), 2018 WL 5905879, at *2 ("To reverse the application of the First Amendment—that is, to make online platforms no longer *protected* by the First Amendment but instead *bound* by it as if they were government entities—would undermine Internet users' interests." (emphasis in original)).

270. *See* Goldman & Miers, *supra* note 268, at 195–96.

of private speech regulation. As previously outlined, this project aims at bringing about accountability in content moderation through transparency, due process, reasoned decision-making, participation, etc.[271] Without an agenda for substantive demands, speech protectionist advocates have channeled their efforts into this type of initiative. This focus on procedural fairness has led to successful outcomes but has been ineffective at countering the push for stricter regulation from other actors.

Although shared across the political spectrum, these procedural safeguards have been spearheaded by speech protectionists. The Santa Clara Principles on Transparency and Accountability in Content Moderation are one clear expression of the endeavor. When they launched in 2018, the principles focused on transparency and appeal processes to increase platforms' accountability to their users.[272]

Advocacy organizations and academics "who support the right to free expression online" designed the Santa Clara Principles after considering "how best to obtain meaningful transparency and accountability around internet platforms' increasingly aggressive moderation of user-generated content."[273] The three original principles are the following: companies should publish the numbers of posts removed and accounts suspended; companies should provide notice to each user whose content is taken down or account is suspended about the reason for the decision; and companies should provide an opportunity for timely appeal.[274] Notably, the effort is led by staunch free speech advocacy organizations: the Electronic Frontier Foundation, the Center for Democracy & Technology, the ACLU Foundation of Northern California, and the New America's Open Technology Institute.[275]

The GNI and the RDR Corporate Accountability Index are other important examples of initiatives that I classify under the administrative law project, as they envision a freedom of expression project that focuses on institutional design rather than on substantive rules. The GNI launched in 2008 as a multi-stakeholder effort to help companies navigate government pressure to remove content.[276] The GNI Principles offer guidance to companies on how to conduct their business. Regarding their own internal rules, companies should be transparent about data collection and

271. Bloch-Wehba, *supra* note 13, at 71.

272. SANTA CLARA PRINCIPLES, https://santaclaraprinciples.org/scp1/ [https://perma.cc /H3YB-CDCE] (last visited May 29, 2022). The principles were updated on December 8, 2021. The new version develops the original principles and adds principles about automated moderation and cultural competence.

273. *Id.*

274. *Id.*

275. *Id.*

276. Colin M. Maclay, *Protecting Privacy and Expression Online: Can the Global Network Initiative Embrace the Character of the Net?*, *in* ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS, AND RULE IN CYBERSPACE 87, 87 (Ronald Deibert, John Palfrey, Fafal Rohozinski & Jonathan L. Zittrain, eds., 2010) (discussing the origins of the GNI).

policies to remove content, as well as provide notice to users when their content has been removed.[277] In a similar fashion, the RDR Index does not require specific content moderation rules and instead focuses on procedural guidelines. The recommendations for companies include stakeholder engagement and other mechanisms to ensure robust governance, maximize transparency, provide notice to users whose content is taken down, and report on the impact of their own policies.[278]

The administrative law project has had many limited but important victories. In 2017, appeals on Facebook were only available to people whose profile, group, or page had been removed, but not for individual posts.[279] Now all decisions can be appealed.[280] Facebook made its content moderation rules public as recently as 2018.[281] Through their blogs, companies often share their research on the impact of certain rules or features.[282]

Calls for transparency have often lacked a proposal for an accountability structure. These calls have usually assumed that a diffuse group of actors, including researchers and advocates, will use this information to hold companies accountable.[283] However, accountability requires substantive standards to measure transparency, which so far have been more actively developed by speech regulationists. Therefore, the administrative law project has not only been ineffective in countering the lobby for more aggressive speech restrictions, it may have also aided those lobbying efforts. Notably, when asked about which advocacy efforts have been successful in recent years, speech regulationists who participated in this research point to new prohibitions over harmful speech, whereas speech protectionists answer "more transparency."[284]

277. *The GNI Principles*, GLOB. NETWORK INITIATIVE, https://globalnetworkinitiative.org/gni-principles/ [https://perma.cc/7B6B-JATA] (last visited July 14, 2022); *Implementation Guidelines*, GLOB. NETWORK INITIATIVE, https://globalnetworkinitiative.org/implementation-guidelines/ [https://perma.cc/X9KC-FU7W] (last visited July 14, 2022).

278. *2019 RDR Corporate Accountability Index*, RANKING DIGIT. RTS. 1, 53 (2019), https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf [https://perma.cc/ZK4L-C6PW].

279. Angwin & Grassegger, *supra* note 87.

280. Bickert, *supra* note 83.

281. *Id.*

282. *See, e.g.*, Vijaya Gadde & Kayvon Beykpour, *An Update on Our Work Around the 2020 US Elections*, TWITTER BLOG (Nov. 12, 2020), https://blog.twitter.com/en_us/topics/company/2020/2020-election-update [https://perma.cc/S4SH-DCYH].

283. Susan Ness, *Platform Regulation Should Focus on Transparency, Not Content*, SLATE (Dec. 2, 2020, 4:45 PM), https://slate.com/technology/2020/12/platform-regulation-european-commission-transparency.html [https://perma.cc/VE9J-Y3BH] ("Mandating transparency would increase public pressure on platforms to improve users' online experience.").

284. Interview #01, *supra* note 113; Interview #02, *supra* note 119 (both answering "transparency" when asked about a successful ask from their organizations to companies).

3. *Human Rights*

The human rights project proposes that social media platforms should moderate content according to international human rights law (IHRL). The project became mainstream in 2018, when the U.N. Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, David Kaye, proposed that platforms moderate using "[h]uman rights by default."[285] Although many treaties are relevant for speech regulation, the compass is Article 19 of the International Covenant on Civil and Political Rights, which sets the requirements for legitimate governmental restrictions on freedom of expression.[286] This is the most forcefully articulated demand for platforms to allow more legal content than they currently do.

In terms of practical possibilities for implementation, the human rights project is promising. IHRL has a strong appeal as a legitimate framework.[287] Its proponents invoke the overwhelming adherence of states to U.N. treaties to frame this proposal as putting the consent of states—and their people—back at the forefront of speech regulation, instead of the triumph of a normative preference.[288]

In addition, the project has been coupled with the launch of the Facebook Oversight Board.[289] The Board has chosen IHRL as the framework to justify its decisions, most of which have been more speech protective than Facebook's Community Standards. The Board has been praised for its potential to improve transparency and public reasoning. A less discussed consequence (beneficial or detrimental depending on one's normative preferences) may be that it is well-placed to counter the pressures that have driven companies toward restrictive rules, even though that was one of its original purposes.[290] Indeed, one could think that the Over-

285. *SR Report 2018*, *supra* note 63, ¶ 45.

286. *See* G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights, art. 19 (Dec. 16, 1966).

287. Evelyn Douek, *The Limits of International Law in Content Moderation*, 6 U.C. IRVINE J. INT'L, TRANSNAT'L & COMPAR. L. 37, 44 (2021).

288. Aswad, *supra* note 140, at 45 ("It is only by citing to universal standards embodied in international human rights law that Twitter can claim to ground its worldwide rules in a fair manner."). For a critical perspective, see Brenda Dvoskin, *International Human Rights Law Is Not Enough to Fix Content Moderation's Legitimacy Crisis*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y (Sept. 16, 2020), https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd [https://perma.cc/CS53-FLXV].

289. Nick Clegg, *Welcoming the Oversight Board*, META NEWSROOM (May 6, 2020), https://about.fb.com/news/2020/05/welcoming-the-oversight-board/ [https://perma.cc/F23E -2WRA].

290. Interview with Noah Feldman, of Harv. L. Sch. (May 13, 2020) ("One of the factors that was very much in my mind when I first thought of the idea of the Oversight Board was the existence of institutional entities that lobbied Facebook for content takedowns . . . . So I had the thought that what they needed was an independent mechanism . . . . In a sense, the idea of a constitutional court interpreting a constitutional commitment to free expression functions to create an in-

sight Board can be an effective counterweight to Facebook because it is further removed from commercial or business considerations. However, the most relevant difference in perspective might come from the fact that it follows different legitimizing strategies: while Facebook is mostly concerned with satisfying the loudest normative demands,[291] the Oversight Board has chosen to refer to IHRL to justify its decisions.

The project is attractive to U.S. speech protectionists for at least two reasons. First, despite a wide perception that the United States is an outlier in speech regulation (especially around hate speech), it is much more closely aligned to the United Nations' standards than, for example, the European system of human rights is.[292] On that point, in its decision reviewing Facebook's action to restrict former President Trump's access to his accounts, the Oversight Board offered comfort to those concerned that domestic free speech issues are now governed by international norms: "the Board notes that in many relevant respects the principles of freedom of expression reflected in the First Amendment are similar or analogous to the principles of freedom of expression in ICCPR Article 19."[293]

Holocaust denial illustrates how IHRL could bring content moderation rules closer to First Amendment doctrine. As noted above, YouTube and Facebook banned Holocaust denial from their platforms after years of controversy, further diverging from First Amendment standards.[294] Such rules seem incompatible with international human rights law. Although in 1996 the Human Rights Committee held that a French conviction of a Holocaust denier did not violate the Covenant, in 2011 it stated that "[l]aws that penalize the expression of opinions about historical facts are incompatible with . . . respect for freedom of opinion and expression."[295] Accordingly, Kaye, in his report to the Committee, used laws prohibiting such speech as examples of violations of international law.[296] The Over-

---

stitutionalized barrier to asymmetric lobbying for content removal for states, for government actors."). Feldman led the design of the Facebook Oversight Board. *See generally* Feldman, *supra* note 149.

291. Bickert, *supra* note 10, at 266 ("Simply put, there are business reasons that a big social media company must pay attention to what the world thinks of its speech rules.").

292. *See, e.g.*, Nadine Strossen, *United Nations Free Speech Standards as the Global Benchmark for Online Platforms' Hate Speech Policies*, 29 MICH. ST. INT'L L. REV. 307, 333–34 (2021); Sarah H. Cleveland, *Hate Speech at Home and Abroad*, *in* THE FREE SPEECH CENTURY 210, 230–31 (Lee C. Bollinger & Geoffrey Stone eds., 2018); Evelyn Mary Aswad, *To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?*, 77 WASH. & LEE L. REV. 609, 624 (2020).

293. *Case decision 2021-001-FB-FBR*, OVERSIGHT BD. (May 5, 2021), https://oversightboard.com/decision/FB-691QAMHJ/ [https://perma.cc/4MN4-3TJ2].

294. Bickert, *supra* note 87; *Our Ongoing Work to Tackle Hate*, YOUTUBE OFF. BLOG (June 5, 2019), https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate?m=1 [https://perma.cc/Q8LC-8E9K].

295. U.N. Hum. Rights Comm., General Comment No. 34, ¶ 49, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011).

296. David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the*

sight Board has also noted that "laws establishing general prohibitions of expressions with incorrect opinions or interpretations of historical facts, often justified through references to hate speech, are incompatible with Article 19 of the ICCPR, unless they amount to incitement of hostility, discrimination or violence under Article 20 of the ICCPR," citing the Human Rights Committee's General Comment 34 and the U.N. Special Rapporteur on freedom of expression's report.[297]

Admittedly, the system of regional and international human rights also includes less protective standards,[298] but those promoting the project have amplified the aspects of IHRL that are most protective of speech and dovetail with First Amendment jurisprudence.[299] The global organization Article 19 analyzed the gaps between IHRL and platforms' rules on hate speech, "terrorist" and "extremist" content, morality-based restrictions, and "fake news," finding that corporate rules reach protected speech.[300] Evelyn Aswad, now a member of the Facebook Oversight Board, has similarly argued that Twitter's rules on hateful speech are too broad and thus incompatible with Articles 19 and 20 of the ICCPR.[301]

Second, unlike the state action requirement of the First Amendment, the human rights apparatus includes the United Nation's Guiding Principles on Business and Human Rights (UNGPs). Scholars have read this instrument as setting a soft legal duty for companies to follow IHRL.[302] Being able to invoke a soft-law duty to protect speech has surprisingly opened up a world of lobbying possibilities.

Take the example of how former ACLU President Nadine Strossen explains the turn to IHRL. She frames the problem: "[p]latforms have become the most important forums for the exchange of information and ideas" but "the [p]latforms' content moderation policies are not constrained by the First Amendment's free speech guarantee."[303] Therefore, "[a]lternative tools are required."[304] The tool she promotes is Kaye's proposal to urge platforms to adopt IHRL "[i]nvoking the U.N. 2011 Guiding Principles on Business and Human Rights."[305]

Promotion and Protection of the Right to Freedom of Opinion and Expression, ¶ 22, U.N. Doc. A/74/486 (Oct. 9, 2019).

297. *Case decision 2021-005-FB-UA*, OVERSIGHT BD. (May 20, 2021), https://oversightboard.com/decision/FB-RZL57QHJ/ [https://perma.cc/8GED-L9HS].

298. *See generally* Amal Clooney & Philippa Webb, *The Right to Insult in International Law*, 48 COLUM. HUM. RTS. L. REV. 1 (2017) (arguing that international law insufficiently protects insulting speech through inconsistent norms and case law).

299. Aswad, *supra* note 140, at 637 (2020); Strossen, *supra* note 291, at 317–18.

300. *Side-stepping Rights: Regulating Speech by Contract*, ARTICLE 19 1, 18–34 (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB.pdf [https://perma.cc/UB9D-R6F4].

301. Aswad, *supra* note 140, at 46–47.

302. *Id.* at 39.

303. Strossen, *supra* note 292, at 307.

304. *Id.* at 308.

305. *Id.*

Why is this semi-legal expectation that companies will follow standards designed for states relevant in advocacy work done outside courts or institutions bound by law? Strossen's project urges companies to voluntarily adopt these principles that, in her own terms, "dovetail with major U.S. free speech principles."[306] Surprisingly, while this difference is understandably significant as a legal matter, it has also affected how advocates and scholars imagine companies' responsibility or moral duties. Not only do advocates gravitate toward the frameworks that coincide with their interests (as it is now part of conventional wisdom), but also legal texts exercise power over what advocates imagine as possible normative projects.

The IHRL project faces a similar challenge as the First Amendment project: it allows too much speech—even for most of those who would prefer more protective rules. Once again, most scholars and the general public do not think that allowing all legal speech on platforms is the best alternative. However, once IHRL is established as an appropriate normative framework, exceptions become possible.

The Center for Democracy & Technology does not generally make public recommendations about what rules private companies should adopt. In the Facebook Oversight Board's case reviewing Facebook's decision to suspend former President Trump's access to his accounts, the organization submitted a comment (similar to an amicus curiae before the Oversight Board). The submission begins by stating the international human rights standard that would apply to the case: "There is a strong presumption against the validity of prior censorship in international human rights law."[307] However, the submission proceeds, "there are important differences between a state actor" and an online service.[308] The comment goes on to articulate a standard that Facebook could adopt to regulate the suspension of accounts of political figures explicitly referencing the Rabat Plan of Action, a set of recommendations launched by the U.N. Office of the High Commissioner for Human Rights.[309] This example illustrates how adopting IHRL as a possible framework allows both for endorsing some of its elements and creating exceptions to other areas; ultimately, it enables the articulation of a normative proposal.

In sum, once IHRL emerges as a possible authoritative set of standards to anchor normative preferences, the project can grow into a substantive agenda comprised of some of its standards, new formulations of those standards, and even new exceptions. The project is incipient, but IHRL and its variations may open the door for speech protectionists to

306. *Id.* at 333.

307. Emma Llansó, *CDT's comments to Facebook Oversight Board on 2021-001-FB-FBR (Case Regarding Suspension of Trump's Account)*, CDT BLOG (Feb. 11, 2021), https://cdt.org/insights/cdts-comments-to-facebook-oversight-board-on-2021-001-fb-fbr-case-regarding-suspension-of-trumps-account/ [https://perma.cc/M57X-B2WL].

308. *Id.*

309. *See id.*

build a normative agenda that does not allow all legal speech but resists regulationists' calls for stricter rules. The most surprising aspect of this development is how the difference in legal texts (compare the First Amendment's state action requirement to the UNGPs' expectations for private actors) shape normative expectations even when the law is by no means binding.

IV. THE EFFECTS OF LAW IN ONLINE SPEECH GOVERNANCE

Calls to include civil society in governance make two assumptions. First, participation is presented as a procedural safeguard agnostic concerning any normative preference. In this sense, civil society participation is assumed to be a neutral device: it offers all viewpoints a chance to influence policy.[310] Second, civil society participation is part of a private rulemaking process that happens beyond the law. The legal system may prohibit some very limited categories of harmful speech and it reserves the negotiation of online expression rules to other actors. The account I have offered here challenges both assumptions.

Law is at work on at least three levels. On the institutional level, law has set up the terrain on which online speech governance takes place.[311] Regulatory decisions (and lack thereof) have enabled a highly concentrated media environment. Lina Khan, Tim Wu, and others document how antitrust law has evolved to define harm exclusively in reference to consumer welfare, usually measured by short-term pricing.[312] This interpretation—and corresponding enforcement decisions—has not captured adequately the market structure of informational capitalism: social media companies often offer their services for free.[313]

310. *See, e.g.*, *Meta's stakeholders*, META, https://transparency.fb.com/policies/improving/our-stakeholders [https://perma.cc/M4MN-CPRF] (Jan. 19, 2022) ("In a sense, we consider our stakeholders to be every person or organization that may be impacted by the policies we set out in the Facebook Community Standards and Instagram Community Guidelines, which apply to every post, photo and video shared on Facebook or Instagram. Of course, because we can't meaningfully engage with billions of people, we seek out organizations that represent the interests of others . . . .").

311. *See generally* Amy Kapczynski, *the Law of Informational Capitalism*, 129 YALE L.J. 1460 (2020); Yochai Benkler, *The Role of Technology in Political Economy: Part I*, L. & POL. ECON. BLOG (July 25, 2018), https://lpeproject.org/blog/the-role-of-technology-in-political-economy-part-1/ [https://perma.cc/86Y6-2WDZ] (challenging the idea that technology evolves in an institutional vacuum).

312. *See* Lina M. Khan, *Amazon's Antitrust Paradox*, 126 YALE L.J. 710, 721 (2017); TIM WU, THE CURSE OF BIGNESS: ANTITRUST IN THE NEW GILDED AGE 10 (2018); K. SABEEL RAHMAN, DEMOCRACY AGAINST DOMINATION 72 (2017); Dina Srinivasan, *The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy*, 16 BERKLEY BUS. L.J. 39, 82 (2019).

313. Khan, *supra* note 312, at 792–93 (showing how an exclusive focus on pricing kept the antitrust agencies from sufficiently scrutinizing Facebook's purchase of WhatsApp and Instagram).

Law has constructed private power not only by not breaking up bigness but also by encouraging its creation. Julie Cohen argues that law constructed private power by providing very low protection to user data and thus allowing firms to appropriate it very cheaply.[314] Besides enabling the accumulation of resources, law also isolates it from competition by affording high protections to firms' trove of data and intellectual property.[315] This double regime creates a "first-come, first-served" environment for obtaining data flows and protections against claims to the resource pool.[316]

The high concentration of power in a handful of firms is an essential premise of civil society participation since civil society has the capacity to participate in only a handful of governing structures.[317] Further, lobbying companies is only attractive if the companies control platforms of significant importance for public discourse. If there were many platforms of equal importance, lobbying would be too expensive and probably ineffective. In fact, the history of the media in the United States shows that pressure from civil society groups to restrict speech has worked best in other highly concentrated industries, although its close examination is beyond the scope of this Article.[318]

On the second level, law has produced in these powerful firms a willing partner for pressure groups. The legal conditions that have enabled their bigness (low protection for users' data and high protections for firms' intellectual property) are interests to protect. In addition, Section 230 of the Communications Decency Act offers a very favorable regulatory regime in exchange for companies being good Samaritans.[319] Overall, the current legal environment is greatly favorable to social media giants. Stalling unfavorable reforms provides incentives to accommodate the preferences of those who can put their good Samaritan's name at risk. Pleasing civil society's and government officials' speech moderation preferences is a low-cost concession to appease those who could push unfavorable reforms directly or indirectly. It is in this context that speech regulationists have found in these firms a site for advocacy more promising than U.S. governmental institutions. The "new governors" are both capable and more willing than government actors to respond to the interests of speech regulationists.

---

314. COHEN, *supra* note 148, at 51. *See generally* Kapczynski, *supra* note 311 (reviewing Cohen's book).

315. COHEN, *supra* note 148, at 45.

316. *Id.* at 52.

317. Interview #28, *supra* note 123 (explaining the organization's decision not to respond to companies' solicitation for input to prioritize other activities).

318. *See e.g.*, PAUL STARR, THE CREATION OF THE MEDIA: POLITICAL ORIGINS OF MODERN COMMUNICATIONS 295–326 (2004) (reviewing the influence of lobbying groups in the movie industry).

319. Klonick, *supra* note 1, at 1605 (describing the origins of Section 230 of the Communications Decency Act and how this law was passed "in order to incentivize and protect intermediaries' Good Samaritan blocking of offensive material").

Finally, at the level of ideology, law has shaped the agendas of the advocates who negotiate with companies. Technological innovation may have driven a cultural change in the social views over freedom of expression. Looking at how advocates have engaged legal discourse and presented their normative views to the public adds a new dimension to explain that cultural change.

On one hand, advocates have picked the frameworks that are more promising for reaching their preferred normative outcomes. A clear illustration of this strategy is the regulationists' dual relationship to platform power. Despite their acknowledgment of platforms' power, regulationists have strongly resisted efforts to view platforms as public forums which would extend the reach of the First Amendment's laissez faire approach to speech regulation.

On the other hand, legal frameworks affect how advocates conceive their normative preferences.[320] We are perhaps used to thinking of normative preferences as prior to the legal framing: depending on one's interests, one may construct legal arguments in one way or another. In this case, however, legal texts and available legal frameworks may have influenced how some advocates imagine normative options.

A comparison of speech protection engagement with the First Amendment and with IHRL is an interesting vantage for locating the influence of legal texts over advocates' imagination. The First Amendment's state action requirement has given fertile ground to regulationists to mobilize alternative legal frameworks to reach constitutionally protected speech. On the contrary, it has been paralyzing for speech protectionists. This prerequisite has left speech protectionists without a meaningful vision for private actors.

The UNGPs, on the other hand, have been interpreted as an expectation that companies will follow IHRL in their content policies. This element is expanding the imagination of speech protectionists over how to bridge the public–private divide in ways not available when thinking through the U.S. constitutional rights framing. The possibility of anchoring a substantive agenda in terms of legal duties or expectations is opening up a new space to articulate normative demands.

The exact influence of each group is difficult, perhaps impossible, to measure. Content moderation rules have become stricter, but it is unknown how much more stringent they would have become if protectionists had not countered the calls for stricter regulation. Another limitation is the difficulty of disentangling the pressures from governments, advertisers and regulationist civil society. This Article hypothesizes that advocates' engagement with legal discourse has produced asymmetric lobbying agendas within U.S. civil society. Because of this intellectual asymmetry, regulationists might be more effective risk makers. Corporate interest in managing

320. Kapczynski, *supra* note 21, at 876 (describing how social movements' interests do not preexist law but are constituted through it).

reputational risks can explain how this intellectual asymmetry might have contributed to the normative evolution of content moderation.

What do these distributional implications mean for the project of civil society participation as a democratizing tool? The fact that the project has had these distributional implications does not mean we should reject it. One possible answer is that this is exactly what democratic mechanisms are about. After all, democracy is not about neutrality, but about channeling political projects.

The issue speaks to the broader question about what it means to democratize online speech governance while resisting the traditional or immediate answer of resorting to existing democratic institutions. A response in projects critical of both private power and governmental democracy is to identify or open spaces for contestation and participation.[321] The paradox of contestation-participation is that, when formalized, it risks empowering the actors better organized to make use of those institutions when these actors might not adequately represent the interests of the boarder society. Once contestation space becomes institutionalized, these spaces can replicate extra-institutional power relations and be subject to capture.[322]

The path forward should involve recursive diagnosis and reform. If the goal is to build points of entry so that people may have an opportunity to participate in the decisions that affect society, one can aim at constantly evaluating the voices that are disadvantaged by existing institutional arrangements and identify how they can be included. The core purpose of this Article has been to conduct this type of evaluation. Part V identifies some points for reform and intervention.

V. CIVIL SOCIETY PARTICIPATION AS A DEMOCRATIZING MECHANISM

Promoting civil society participation as a democratizing tool is a risky project premised on the existence of a highly concentrated or coordinated industry. Not only does it leave corporate power unchallenged, in important ways, it can reinforce it: civil society and companies become partners in their efforts to govern. These conditions may be unacceptable for those who believe that the main problem with social media is just how powerful these companies are.[323] Participation also requires the existence of a robust civil society complex. The funding and legitimacy of these organizations poses significant challenges that this Article does not address.[324]

321. *See, e.g.*, Sheila Jasanoff, *Constitutional Moments in Governing Science and Technology*, 17 SCI. ENG. ETHICS 621, 623–24 (2011).

322. Ganesh Sitaraman, *The Puzzling Absence of Economic Power in Constitutional Theory*, 101 CORNELL L. REV. 1445, 1500 (2016); Nico Krisch, *The Pluralism of Global Administrative Law*, 17 EUR. J. INT'L L. 247, 276 (2006).

323. WU, *supra* note 312, at 11 (arguing that the existence of companies as powerful as Facebook or Google threatens democracy).

324. *See generally* THE REVOLUTION WILL NOT BE FUNDED: BEYOND THE NON-PROFIT INDUSTRIAL COMPLEX (INCITE! ed. 2007); Emanuel Moss, Elizabeth Anne

Beyond funding, civil society needs a high level of sophistication and support from the companies to fully comprehend and engage meaningfully in governance discussions.[325]

But if one acknowledges that the public should have a say in how speech is governed in those spaces that matter most to them and that democratic institutions are not well placed to occupy that role, then civil society participation may still be an attractive option. Three actors can improve the existing participatory channels: companies can improve the institutional design; civil society actors can make better use of this mechanism; and the state can set up better conditions for these conversations to take place.

On how to improve civil society engagement, the low-hanging fruit is to include the groups that are not often consulted. A clear example are organizations of sex workers. Geopolitical imbalances are still notorious.[326] Even in the cases of global participatory efforts such as the Twitter's Trust and Safety Council and Facebook's stakeholder engagement team, organizations in the global South report weaker ties than organizations in central countries.[327]

Stakeholder engagement consultation often lacks the "notice" part of notice-and-comment rule making.[328] Open consultations could result in less meaningful engagement between the company and each participating advocate. However, giving notice to the public about major policy changes would take away some of the gatekeeping power that companies currently have to recognize the groups that should participate.[329] In this sense, the Facebook Oversight Board recently recommended Facebook "ensure meaningful stakeholder engagement" on proposed changes to its Dangerous Individuals and Organizations policy "including through a *public call for inputs*."[330] A mixture of open calls and solicited feedback with opportunities for deeper engagements could be a better balance.

Watkins, Ranjit Singh, Madeleine Clare Elish & Jacob Metcalf, *Assembling Accountability: Algorithmic Impact Assessment for the Public Interest*, DATA & SOC'Y 9, 22 (2021), https://datasociety.net/wp-content/uploads/2021/06/Assembling-Accountability.pdf [https://perma.cc/897Z-Q5DC] [hereinafter *Assembling Accountability*].

325. *Assembling Accountability*, *supra* note 324, at 20.

326. *SR Report 2018*, *supra* note 63, ¶ 54.

327. For example, in interviews conducted with organizations based in Latin America, participants overwhelmingly report engaging companies through their trusted partner programs, which have the primary goal of providing local context for the enforcement of policies in specific cases rather than providing advice as part of policy development processes. *See id.*

328. 5 U.S.C. § 553(b) ("General notice of proposed rule making shall be published in the Federal Register . . . .").

329. Richard B. Stewart, *Remedying Disregard in Global Regulatory Governance: Accountability, Participation, and Responsiveness*, 108 AM. J. INT'L L. 211, 263 (2014).

330. *Case decision 2021-006-IG-UA*, OVERSIGHT BD. (July 8, 2021), https://www.oversightboard.com/decision/IG-I9DP23IB/ [https://perma.cc/BA65-765C] (emphasis added).

Transparency also needs to apply to stakeholders. Companies and civil society sometimes prefer not to list publicly who participates in each consultation. First, some organizations may face safety concerns, especially outside the United States. These reasons may be serious and should be accommodated, but also exceptional. Second, civil society organizations may not want their names used as a legitimatizing stamp on policy announcements.[331] The costs of keeping the names of participants in the dark, however, outweigh that consideration. Opacity enables companies to divide and conquer civil society, which makes it very hard for civil society organizations to build alliances or understand who they are arguing against.[332] More importantly, if civil society engagement is a means to represent the public interest, it is important to make the process accountable to the general public, not just to the actors actively involved. To that end, excluded associations can hold participating actors accountable. Because informal conversations are so common, this would open up some parts of the process with the risk of shifting more of the process to less institutionalized engagements.[333]

How advice is processed also needs to be much more transparent.[334] The lack of feedback is bad for companies and for civil society. A statement of how feedback was processed gives grounds for civil society to develop advocacy strategies.[335] For firms, it can curb concerns that the main reason behind rejecting bad suggestions is that they would harm revenue.[336] Most importantly, engagement needs to be ongoing throughout policy development and enforcement instead of a one-off conversation.

Finally, more aspects of content moderation need to be subject to stakeholder engagement. Stakeholder engagement, as a risk management tool, is mostly focused on the elements of content moderation that are

331. *See* Matsakis, *supra* note 125; Interview with participant #38, head of an organization in Latin America, (June 11, 2020) (stating that she is not sure whether the organization is publicly listed as a member of the Twitter Trust and Safety Council and she could not remember the organization being contacted for policy advice). The organization is listed as part of the Twitter Council.

332. *See generally* Robert O. Keohane & Joseph S. Nye, Jr., *The Club Model of Multilateral Cooperation and Problems of Democratic Legitimacy*, *in* EFFICIENCY, EQUITY, AND LEGITIMACY: THE MULTILATERAL TRADING SYSTEM AT THE MILLENNIUM EFFICIENCY 271 (Roger Porter, Pierre Sauvé, Arvind Subramanian & Americo Beviglia Zampetti eds., 2001) (describing civil society's work in transnational coalitions).

333. Cass R. Sunstein, *Output Transparency vs. Input Transparency*, (Aug. 19, 2016) (unpublished manuscript) (available on SSRN), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2826009 [https://perma.cc/E9BT-WV9M] (discussing the benefits and risks of internal transparency in government).

334. The standard practice of administrative agencies to publish a summary of comments and how they were considered can serve as a rudimentary model.

335. Interview with participant #14 (May 21, 2020) (explaining that they would be able to provide better arguments when engaging with companies if they had a better sense of how previous feedback has been considered).

336. *See, e.g.*, Interview #29, *supra* note 127 (expressing concerns that their recommendation to shut off the platform seventy-two hours before a presidential election has been rejected because it would undermine profit).

more transparent and therefore more immediately subject to public criticism. But rules are a tiny fraction of the content moderation complex, which includes algorithmic distribution of content, manual and automated enforcement of rules, partnerships with fact-checkers, and collaboration among social media companies and with other firms such as security firms and payment service providers.[337] Transparency is essential to civil society participation because civil society can only comment on what it can see.[338]

But transparency is riskier than often assumed. Increasing transparency can entrench the power of those who are organized and resourced to leverage that information to advance their interests.[339] Calls for transparency must be backed up by a structure of actors capable of benchmarking transparency reports against substantive standards. Funding and capacity for groups to represent their interests are crucial dimensions that this Article does not explore. The dimension explored here is intellectual: If speech protectionists do not develop such standards, increased transparency is likely to reinforce the imbalance of power between advocacy groups.

Indeed, institutions are not enough. For institutions to enhance civic participation, there needs to be not just better mechanisms of engagement but also actors who make full use of them.[340] For speech protectionists, the state-market divide has been paralyzing; the current market cannot provide the pluralist environment that they wish for. A way out of the current deadlock could come from replacing an ideal of innovation and competition with an ideal of pluralism. Imagining the conditions for a robust public sphere demands we move away from the marketplace of ideas and turn toward at pluralist environment with both space for disagreement and for trusted information. Engaging a new ideal could be deeply generative and take lobbying efforts in new directions. This new ideal could also lead speech protectionists to turn to the state as a free speech ally.

Building the conditions for a pluralist speech environment—one where different spaces serve different expressive needs—looks very different from leaving it to the market to provide it. As a starting point, it may allow for a normative view of what global platforms should look like. If these platforms serve a specific role within a larger pluralist environment, it may become possible to design an agenda tailored to that specific role. Perhaps "corporate IHRL" develops into a set of rules that protects speech on platforms while allowing for adaptations to platforms' architecture and

337. Douek, *supra* note 13, at 5.

338. Bloch-Wehba, *supra* note 13, at 76 ("Full participation by the public and civil society is, indeed, *contingent* on transparency . . . .").

339. *See* David E. Pozen, *Transparency's Ideological Drift*, 128 YALE L.J. 100, 157 (2018).

340. K. Sabeel Rahman, *Policymaking as Power-Building*, 27 S. CAL. INTERDISC. L.J. 315, 353 (2018).

giving leeway to certain corporate decisions. As explored earlier, having IHRL as a framework from which to draw standards and create exceptions may be necessary for speech protectionists to develop a normative project.

The marketplace of ideas assumes that chaos leads to truth. Instead, free speech advocates might look at the rules and circumstances necessary for trusted journalism and the production of knowledge. Beyond platforms, it might be relevant to engage other actors, think through their roles in the public sphere, and whether fixes are required.

Pursuing a pluralist environment without just relying on the market to provide it may also open doors to turn to the state. The state has, indeed, at least three important roles to play. First, as extensively argued, it can mandate transparency and even civil society participation.[341] Second, antitrust is often invoked as a promising tool. However, a pluralist speech environment is not one in which big platforms are broken up into smaller versions of themselves. It is one that supports diverse spaces for communication. A better public sphere does not need more market, but more trusted institutions and publicly funded alternatives that escape the social media business model. The state can support and subsidize media actors and build public architecture to create a broader ecosystem.[342] The state has the funding capacity to help develop the spaces and actors that speech pluralists may identify as missing. Third, within that broader ecosystem, giant platforms may occupy a better-defined role. If platforms are to provide more speech protection than other spaces, emerging normative projects for the state may gain more traction (such as partial must-carry laws or a more active role for courts and legislatures in ensuring that companies do not unduly inhibit the public sphere).[343]

None of these projects will be without challenges. But they may lead toward a more representative conversation about the role of giant platforms and toward the vibrant speech environment that the current market cannot provide.

CONCLUSION

What Jonathan Zittrain calls the process era of digital governance is the contemporary desperate attempt to legitimize online speech governance circumventing the political discussions about the substance of con-

341. Besides the calls for voluntary transparency discussed in Section III.B, scholars have also proposed and discussed legal transparency mandates. *See generally* FRANK PASQUALE, THE BLACK BOX SOCIETY: THE SECRET ALGORITHMS THAT CONTROL MONEY AND INFORMATION (2015); Daphne Keller, *Some Humility about Transparency*, CTR. FOR INTERNET & SOC'Y (Mar. 19, 2021, 3:09 AM), http://cyberlaw.stanford.edu/blog/2021/03/some-humility-about-transparency [https://perma.cc/Y82N-BDQ7]; MARTHA MINOW, SAVING THE NEWS: WHY THE CONSTITUTION CALLS FOR GOVERNMENT ACTION TO PRESERVE FREEDOM OF SPEECH 136 (2021) (calling for governmental support of civil society to protect internet users).

342. MINOW, *supra* note 341, at 102 (arguing that the First Amendment allows and even calls for government action to support the press).

343. Keller, *supra* note 6, at 24; Lakier & Tebbe, *supra* note 256.

tent moderation. The endeavor assumes that because agreeing on what speech rules are preferable is impossible, we should focus instead on procedural safeguards. These projects present themselves thus as agnostic concerning substance: they merely institute devices to channel normative disagreements. But this disengagement from substantive debates is impossible. Seemingly neutral procedural fixes distribute power and favor some normative views over others. Specifically, civil society participation in digital governance through notice-and-comment style procedures may distribute power to the advocacy groups in a better position to make use of them. To understand these dynamics, it is necessary to look at institutional projects not in isolation but in relation to the legal and institutional environment where they are embedded.

If the goal is to democratize online speech governance—meaning to find ways for the public to have meaningful opportunities to influence it—, then we need to engage in a recursive evaluation of the distributional implications of new processes and adjust accordingly so that everyone gets a fighting chance. My hope is that making those implications visible will open up new spaces for normative imagination and civic participation.