ARTICLE

Against Privatized Censorship: Proposals for Responsible Delegation

MOLLY K. LAND*

From beheadings to hate speech, the internet is awash in material that poses risks to a range of state objectives. And in light of recent events—from Facebook's role in the genocide in Myanmar to the ways in which social media was used by the perpetrator of the Christchurch massacre—the question is no longer whether, but how, states will regulate social media platforms. Governments, however, have responded to the problem of harmful online content by privatizing the regulation of speech. Germany, France, Australia, the United Kingdom, and the European Union are joining countries such as China, Turkey, and Thailand in enacting laws that delegate to platforms extensive authority to remove speech on their behalf.

This Article is the first to examine the legality of such privatized censorship. Using international law as a baseline, the Article argues that delegating unconstrained authority to platforms to determine what speech is permitted or prohibited transforms platforms into state actors that must then ensure their decisions comply with human rights norms. It further argues that naked delegations, unaccompanied by safeguards, are unlawful under human rights law. The Article then develops a framework for the lawful regulation of social media platforms. The Article considers proposals for accountability based in both law and code, arguing that regulators must not only establish oversight mechanisms but must also seek changes in platform structure and business models in order to ensure the responsible governance of online speech.

* Catherine Roraback Professor of Law and Human Rights, University of Connecticut School of Law, and Associate Director, Human Rights Institute, molly.land@uconn.edu. I am grateful for feedback received at the ASIL Mid-Year Research Forum 2019; the Harvard Law School International Law Workshop; the University of Connecticut School of Law Faculty Workshop; the Drexel Law School Faculty Workshop; the Business and Human Rights Scholars Association 4th Annual Conference; the 2016 Freedom of Expression Scholars Conference at Yale Law School; the International Law Colloquium of the Center for International and Comparative Law at St. John's University School of Law; and the International Studies Association's Human Rights Conference. Specific thanks to William Alford, Evelyn Aswad, Jack Balkin, Bethany Berger, Gabriella Blum, Rita Cant, Eileen Doherty-Sil, Rebecca Hamilton, Peggy Hicks, Larry Helfer, Mark Janis, Rikke Frank Jørgensen, Michael Karanicolas, Rachel Lopez, Gerald Newman, Federica Nieri, Barrie Sander, Jeremy Sheff, Richard Wilson, and the students in the HLS International Law Workshop for their comments and feedback. I am grateful to Zeynep Aydogan and Camden Weber for excellent research assistance.

I. INTRODUCTION

Responding to harmful content on the internet is an important governmental prerogative. States seek to vindicate a variety of human rights and public policy goals through control of content online, from removing extremist material to remedying the harms of degrading, demeaning, and hateful speech. States have long sought to realize their objectives by asserting authority over critical internet control points—namely, the private companies that develop, own, and operate the infrastructure of speech online.

In recent years, however, we have witnessed a profound shift in these methods for asserting authority. States have moved beyond attempting to *control* private platforms to *deputizing* them—delegating to these private actors the responsibility and authority to police and govern internet content. Through techniques as diverse as legal liability to coerced "self-regulation," governments are shifting authority over the regulation of speech to social media platforms. This shift constitutes not a privatization of the internet, for private actors have long controlled the internet. Rather, it constitutes a privatization of speech regulation. States increasingly rely on private actors to make decisions about who is allowed to speak and, in the process, insulate this exercise of public authority from both national and international accountability mechanisms.

As more and more jurisdictions have or are poised to enact legislation designed to delegate ever greater authority over speech to social media platforms, we must address whether such delegation is lawful. Jurisdictions such as China, Turkey, and Thailand have long required social media companies to police the content that appears on their platforms. Now, countries such as Germany, Australia, France, and the United Kingdom have made or are contemplating similar moves. Despite extensive concern about the power wielded by these "New Governors" of speech[1] and the accountability gaps that this creates,[2] there has yet to be a thorough consideration of the lawfulness of such privatized censorship.[3]

This Article uses international law to evaluate the lawfulness of privatized censorship. It relies on international law as a baseline given the wide variation in national laws regarding speech as well as the global nature

1. Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1599 (2018).

2. Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 69-70 (2019).

3. Content moderation is only "censorship" if it is unlawful. *See* Anupam Chander & Uyên P. Lê, *Data Nationalism*, 64 EMORY L.J. 677, 679-80 (2015); Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2299-300 (2014). This Article uses the term "censorship" to refer to content moderation done pursuant to broad delegated authority because of the disproportionate (and thus unlawful) impact of such delegated authority.

of social media platforms. Under international law, when a state empowers a private actor to exercise governmental authority, the resulting action is treated as the action of the state and therefore must comply with the state's human rights obligations. In this way, states may not insulate themselves from responsibility for their actions that impact freedom of expression by "laundering"[4] state authority through private platforms or leveraging these platforms to target speech that they could not lawfully censor themselves. Further, the Article argues that delegation itself is unlawful under human rights law unless accompanied by meaningful safeguards to ensure accountability. Public-private cooperation to address the harms of online speech is essential, but delegation is only permissible under conditions that ensure transparency and accountability. The Article then develops a set of recommendations for ensuring those conditions are met.

The Article makes two important contributions to the literature on human rights and content moderation. First, it provides a legal analysis of privatized censorship that can serve as a foundation for responding to the deep intertwining of public and private authority that pervades the regulation of speech online.[5] This Article is the first to examine the limits that international law puts on broad delegations of authority over speech.

Second, the Article offers a roadmap for more responsible delegation to social media platforms that navigates the competing demands of innovation, scale, and human rights. Companies, scholars, experts, and activists have long raised concerns about intermediary liability and advocated additional oversight and transparency.[6] This Article surveys the solutions that have been advanced and develops a hybrid model that combines recommendations regarding both the appropriate level of substantive liability as well as procedural mechanisms for promoting accountability. Thus, the Article suggests a graduated scale of liability based on the nature of content regulated, combined with judicial review in contested cases, and technological or design choices designed to augment

4. Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power over Online Speech* 3-10 (Hoover Working Grp. on Nat'l Sec., Tech., & Law, Aegis Series Paper No. 1902, 2019), https://www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech.

5. *See, e.g.*, Bloch-Wehba, *supra* note 2; Molly Shaffer Van Houweling, *Sidewalks, Sewers, and State Action in Cyberspace*, https://cyber.harvard.edu/is02/readings/stateaction-shaffer-van-houweling.html; Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation*, 33 BERK. TECH. L.J. 287 (2018).

6. Although the term "intermediary" can include not only social media platforms but also a range of other entities such as payment or sharing economy intermediaries, this Article focuses on "third-party platforms that mediate between digital content and the humans who contribute and access this content." LAURA DENARDIS, THE GLOBAL WAR FOR INTERNET GOVERNANCE 154 (2014).

user autonomy to provide a greater range of remedies for the harms of online speech.

Part II introduces the problem of delegated censorship. This Part provides a brief overview of the early human rights challenges presented by state regulation of intermediaries and examines how these challenges have shifted over time. This Part tells the story of the rise of privatized censorship through a series of events including the Syrian refugee crisis, Brexit, the 2016 U.S. elections, and the massacre in Christchurch, New Zealand. It argues that the most significant human rights challenges today are not straightforward government demands on platforms to violate human rights law, but rather the wide array of sophisticated and nuanced political and legal techniques employed by governments to appropriate the power of intermediaries in order to censor speech in furtherance of state objectives.

Part III then evaluates delegated censorship under two branches of international law: principles of state responsibility and international human rights law. Under principles of state responsibility, when the state delegates governmental authority to a private actor, the resulting activity is state action that must comply with the state's international obligations. Requiring intermediaries to assess the lawfulness of the speech on their platforms engages them in the act of making law—an activity that must comport with limits on the state's own lawmaking authority.

Next, this Part uses international human rights law to argue that naked delegations of authority to regulate speech violate guarantees of free expression. Because they are making decisions about speech that is not their own, platforms do not have sufficient incentives to limit the burdens they put on speech. Thus, unconstrained delegation of authority to regulate the speech of others will inevitably result in overly broad censorship. This does not mean, however, that delegation is always unlawful. Human rights law does not prohibit delegation, but rather requires states to ensure such delegation is accompanied by safeguards to protect rights.

Part IV then discusses what responsible delegation mechanisms might look like. This Part synthesizes existing proposals for intermediary liability to develop a three-pronged approach of differentiated liability, definitional specificity, and mechanisms for accountability. The proposal advocates for a limited obligation to monitor for child pornography combined with a presumption of immunity that can be rebutted by a company's failure, after notice, to investigate and mitigate harms of disseminating the identified speech. This Part also emphasizes the need for greater specificity in the definitions of harmful speech as applied online, as well as targeted quasi-judicial oversight of the activities of platforms exercising delegated authority to regulate speech.

The governance and accountability problems associated with private control of the internet are not new issues. Leading internet scholars have considered the problem of delegation to intermediaries as a new form of internet regulation,[7] including in the specific contexts of freedom of expression[8] and intellectual property.[9] But the problem today is global. Countries around the world are turning to intermediaries to "solve" the problem of harmful content online. We need international law to ensure these moves will comply with fundamental human rights.

II. THE RISE OF PRIVATE PLATFORMS

Online expression today is largely governed by private companies. Kate Klonick's work illustrates the extent to which these platforms not only own and operate the infrastructure through which vast swaths of speech occurs, but also govern and police this speech as well.[10] Companies such as Google, Facebook, and Twitter moderate our conversations, deliver our news, and keep us connected with acquaintances, friends, and family. These companies now manage our communication and social relationships in ways that can affect everything from our emotions to our sense of dignity, our livelihoods, and even our elections.

It is not only the rise of these "New Governors" that should give us pause. It is also the way in which states are leveraging them to achieve their own ends—thereby creating an end run around rule of law and established systems of checks and balances. States have been extraordinarily adept in devising both legal and technological responses that enable them to assert fairly significant control over the communicative space of the internet. These strategies usually involve controlling the private actors that own and operate the infrastructure and moderate the content on their platforms. This Part describes these historical trends and then focuses on how states have increasingly turned to strategies of privatized censorship.

A. A Short History of State Control

Some of the earliest state efforts to regulate intermediaries were focused on ensuring intermediaries were *protected* from liability. The potential liability of internet intermediaries for publishing, hosting, or

7. Balkin, *supra* note 3, at 2296.

8. Derek E. Bambauer, *Orwell's Armchair*, 79 U. CHI. L. REV. 863, 867-68 (2012).

9. Annemarie Bridy, *Graduated Response and the Turn to Private Ordering in Online Copyright Enforcement*, 89 OR. L. REV. 81, 84-85 (2010); Annemarie Bridy, *ACTA and the Specter of Graduated Response*, 26 AM. U. INT'L L. REV. 559-60 (2011).

10. *See* Klonick, *supra* note 1, at 1662-64.

transmitting harmful content first became a significant issue in the early 1990s. In the United States, early court cases signaled that internet service providers might be liable for content posted by their consumers if they took on an editorial role, raising concerns that intermediaries would as a result be discouraged from moderating or filtering content on their sites.[11]

In reaction, the U.S. Congress passed Section 230 of the Communications Decency Act (CDA) in 1996. Section 230 contains two crucial provisions. Subsection (c)(1) provides, "No provider or user of an interactive computer service shall be treated as the publisher or speaker of any information provided by another information content provider."[12] This subsection thus protects intermediaries from secondary liability for the content they transmit and host. Subsection (c)(2) extends this protection from liability to include affirmative acts undertaken by the intermediary to regulate content. This subsection protects intermediaries from civil liability for "any action voluntarily taken in good faith to restrict access to or availability of material" that might be objectionable as well as actions to make regulatory technology available to others.[13] Although justified today mainly in terms of protecting innovation and freedom of expression,[14] Section 230 played an important role in establishing normative expectations around the appropriateness of intermediaries assuming regulatory roles and developing systems of private control.[15]

Subsequent battles over regulation and the role of intermediaries in controlling online speech were fought over copyright. The internet provided an ideal platform for sharing digitized music. In response to growing copyright infringement, the music industry launched a series of legal and policy initiatives to protect their business models. Copyright owners, dismayed by the rise of peer-to-peer (p2p) music sharing on sites

---

11. ORG. FOR ECON. CO-OPERATION & DEV. [OECD], *The Role of Internet Intermediaries in Advancing Public Policy Objectives*, at 3, DSTI/ICCP(2010)11/FINAL (June 22, 2011), https://www.oecd.org/internet/ieconomy/48685066.pdf; Jeff Kosseff, *Defending Section 230: The Value of Intermediary Liability*, 15 J. TECH. L. & POL'Y 123, 128-32 (2010); Anthony Ciolli, *Chilling Effects: The Communications Decency Act and the Online Marketplace of Ideas*, 63 U. MIAMI L. REV. 137, 147-48 (2008); Mark A. Lemley, *Rationalizing Internet Safe Harbors*, 6 J. TELECOMM. & HIGH TECH. L. 101, 101-02 (2007).

12. 47 U.S.C. § 230(c)(1) (2012).

13. *Id.* § 230(c)(2)(A)-(B).

14. CTR. FOR DEMOCRACY & TECH., SHIELDING THE MESSENGERS: PROTECTING PLATFORMS FOR EXPRESSION AND INNOVATION 5 (2012) [hereinafter SHIELDING THE MESSENGERS]; Ciolli, *supra* note 11, at 148. Anupam Chander argues that protection for internet intermediaries in the United States was instrumental for the success of Silicon Valley, in contrast to other jurisdictions in which intermediaries enjoyed weaker protections. Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639, 642 (2014).

15. Nunziato argues that in creating these expectations, Section 230 was an important step in the privatization of the internet. Dawn C. Nunziato, *The Death of the Public Forum*, 20 BERK. TECH. L.J. 1115 (2005); Dawn C. Nunziato, *Freedom of Expression, Democratic Norms, and Internet Governance*, 52 EMORY L.J. 187 (2003).

such as Napster, directly sued intermediaries for facilitating copyright infringement. Initially limited to instances in which the platform knew of the infringing content,[16] the Supreme Court expanded this approach to hold liable platforms that take active steps to invoke copyright infringing uses.[17] Later, as p2p sites began designing around those legal limitations to avoid secondary liability for copyright infringement, copyright owners started bringing lawsuits against individuals who shared copyrighted material online.[18]

At the same time that they were suing platforms, copyright owners were also lobbying for an international treaty and then national legislation to protect their interests. Both the treaty, the WIPO Performances and Phonograms Treaty, and its implementing legislation, relied on intermediaries to combat copyright infringement. In 1998, Congress passed the Digital Millennium Copyright Act (DMCA) to implement its obligations under the treaty. Section 512 of the DMCA creates conditional immunity from liability for intellectual property violations—a safe harbor—for an intermediary that "responds expeditiously to remove, or disable access to" material claimed to violate intellectual property rights, after being notified of such content.[19] The DMCA's notice and takedown approach was a compromise between the strict liability advocated by copyright holders and the immunity sought by internet service providers.[20] Congress, as well as advocacy groups and internet service providers, worried that strict liability would increase costs, harm the internet service industry, and impede freedom of expression and privacy.[21]

In the late 1990s and early 2000s, states began to seek ever greater control over the online information ecosystem. Two paradigmatic cases emerged, one in France and the other in China. In France, the government sought to require Yahoo! to prevent users from accessing Nazi memorabilia—which was illegal in France—on its platform. The resulting court decision in France upheld this demand and ordered Yahoo! to prevent French users from accessing anti-Semitic material on its site.[22]

16. *See, e.g.*, A&M Records, Inc. v. Napster, Inc. (Napster I), 239 F.3d 1004, 1019, 1022 (9th Cir. 2001) ("Traditionally, 'one who, with knowledge of the infringing activity, induces, causes or materially contributes to the infringing conduct of another, may be held liable as a 'contributory' infringer.'").

17. MGM Studios, Inc. v. Grokster, Ltd., 545 U.S. 913, 940-41 (2005) ("The inducement theory of course requires evidence of actual infringement by recipients of the device.").

18. *See* Ben Depoorter et al., *Copyright Backlash*, 84 S. CAL. L. REV. 1251, 1260-61 (2011) (discussing the turn toward copyright suits against individuals to deter peer-to-peer file sharing).

19. 17 U.S.C. § 512(c)(1)(C) (2018).

20. MILTON L. MUELLER, NETWORKS AND STATES: THE GLOBAL POLITICS OF INTERNET GOVERNANCE 138 (2010).

21. *Id.*

22. *See* Yahoo! Inc. v. La Ligue Contre le Racisme et l'Antisemitisme, 433 F.3d 1199, 1202-04 (9th Cir. 2006).

Although Yahoo! initially sought declaratory relief from courts in California, the controversy was resolved through resort to geolocation technologies. These technologies allowed Yahoo! to prevent access to Nazi memorabilia by French users while allowing this material to remain accessible to users in other jurisdictions.[23]

The Chinese government had early on established comprehensive legal, technical, and social measures to control information online. As part of these efforts, China pressured intermediaries such as Yahoo! and Google to comply with local laws, including laws that violated international human rights. Yahoo!, which had stored user data in China, was ordered to provide the government with data that led to the arrest of several dissidents.[24] There was also public outcry over a 2006 decision by Google to provide censored search services at *google.cn* operating on servers in China.[25] Concern about U.S. intermediaries violating human rights at the direction of the Chinese government and elsewhere prompted Congressional hearings at which executives from several prominent user-facing companies faced questions about their respect for human rights.[26]

By the end of the 2000s, human rights concerns were focused on state orders and demands that violated international human rights law.[27] In authoritarian countries, states engaged in a range of heavy-handed control techniques such as orders to block particular websites or URLs or to terminate individual user accounts.[28] Facing pressure to take actions inconsistent with human rights law in China among other jurisdictions, three prominent Internet companies—Google, Yahoo!, and Microsoft—began conversations in 2006 with a core group of human rights groups and academic institutions that eventually led to the launch in 2008 of the Global Network Initiative ("GNI"), which sought to help companies navigate pressure from governments to take actions that violate human

23. JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 7-9 (2006).

24. Miguel Helft, *Chinese Political Prisoner Sues in U.S. Courts, Saying Yahoo Helped Identify Dissidents*, N.Y. TIMES (Apr. 19, 2007), http://www.nytimes.com/2007/04/19/technology/19yahoo.html?_r=0. These events formed the basis for lawsuits brought against Yahoo! in the United States under the Alien Tort Statute. Anupam Chander, *Trade 2.0*, 34 YALE J. INT'L L. 281, 296-98 (2009).

25. Molly Beutz Land, *Google, China, and Search*, 14(25) ASIL INSIGHT (Aug. 5, 2010), https://www.asil.org/insights/volume/14/issue/25/google-china-and-search.

26. *Id.*

27. Rikke Frank Jørgensen, *Human Rights and Private Actors in the Online Domain*, *in* NEW TECHNOLOGIES FOR HUMAN RIGHTS LAW AND PRACTICE 243, 263 (Molly K. Land & Jay D. Aronson eds., 2018).

28. SHIELDING THE MESSENGERS, *supra* note 14, at 18-20; Jonathan Zittrain & John Palfrey, *Internet Filtering: The Politics and Mechanisms of Control*, *in* ACCESS DENIED: THE PRACTICE AND POLICY OF GLOBAL INTERNET FILTERING 1, 36-38 (Ronald Deibert et al. eds., 2008).

rights norms.[29] The GNI Implementation Guidelines, which provide guidance to companies and form the basis for periodic company assessment, focus their discussion of responsible company decision-making on the efforts the company makes to protect rights when responding to government demands or otherwise complying with local law.[30]

B. New Challenges

The rise of online intermediaries has both amplified and expanded the human rights challenges of public speech on private platforms. Today, states worry not only about Nazi memorabilia, pornography, and copyright, but also fake news, incitement to commit genocide, graphic violence, extremist or terroristic content, hate speech, pro-anorexia or pro-suicide content, harassment, bullying, misinformation, disinformation, and defamation—among many others.[31] The online and offline harms of this content[32] are particularly urgent given the growing dominance and consolidation of social media platforms. Facebook, for example, had 2.5 billion monthly active users as of December 2019.[33] These platforms have a communicative reach that goes far beyond that of any traditional media outlet or government.

The simple dominance of these platforms and the outsized impact that their decisions about content can have on individuals has sparked concern about a range of new human rights problems. Clearly, the question of how companies can and should respond to state requests continues to play a central role in promoting respect for human rights on the internet. Yet the decisions that these companies make about their *own* policies and

29. For a discussion of the founding of GNI, see generally 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 (Ronald Deibert et al. eds., 2010). (Until April 2020, the author participated as an alternate on the GNI Board of Directors on behalf of the Human Rights Institute at the University of Connecticut, which became a member of GNI in 2015. All views expressed here are the author's own.)

30. *See* GLOB. NETWORK INITIATIVE, IMPLEMENTATION GUIDELINES FOR THE PRINCIPLES ON FREEDOM OF EXPRESSION AND PRIVACY (2017).

31. *See, e.g.*, REBECCA MACKINNON ET AL., FOSTERING FREEDOM ONLINE: THE ROLE OF INTERNET INTERMEDIARIES 31-36 (2014) [hereinafter FOSTERING FREEDOM ONLINE] (discussing the varied goals of government regulation of online content).

32. *See* DANIELLE KEATS CITRON, HATE CRIMES IN CYBERSPACE 8-15 (2014); Brian Leiter, *Cleaning Cyber-Cesspools: Google and Free Speech*, *in* THE OFFENSIVE INTERNET: SPEECH, PRIVACY AND REPUTATION 155, 155 (Saul Levmore & Martha C. Nussbaum eds., 2012); DANIEL J. SOLOVE, THE FUTURE OF REPUTATION: GOSSIP, RUMOR, AND PRIVACY ON THE INTERNET 4 (2008).

33. *The Top 20 Valuable Facebook Statistics*, ZEPHORIA DIGITAL MKTG., https://zephoria.com/top-15-valuable-facebook-statistics/ [hereinafter ZEPHORIA DIGITAL MKTG.].

procedures can also have significant consequences for rights.[34] Although the companies themselves have been "reluctant to view content moderation undertaken to enforce their terms of service (TOS) as a human rights issue,"[35] the millions of decisions these platforms make each day about the content on their platforms have significant effects on our ability to generate and share information and expression.[36]

The human rights obligations of platforms is an important question, but one that I address elsewhere.[37] This Article focuses on a different concern—namely, the obligations of states when they rely on intermediaries to police speech. This development is part of a broader shift in which governments are deploying a range of new strategies online—from harassment and misinformation coordinated and boosted through net centers and state-aligned trolls[38] to comprehensive surveillance and tracking[39]—to control not only information but also dissidents and civil society.

A core element of these new strategies is delegation to platforms.[40] James Boyle foresaw this nearly two decades ago, as he argued that states would use privatization to regulate the internet: "[U]nable to respond at Internet speed, and limited by pesky constitutional constraints, the state can use private surrogates to achieve its goals."[41] This is the problem before us now. As Hannah Bloch-Wehba explains: "Rather than simply seeking to enforce domestic law online—whether globally or locally—states are leveraging the infrastructure of private ordering that has long

34. *See generally* TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA (2018).

35. DAVID SULLIVAN, BUSINESS AND DIGITAL RIGHTS: TAKING STOCK OF THE UN GUIDING PRINCIPLES FOR BUSINESS AND HUMAN RIGHTS IN THE ICT SECTOR 16 (2016).

36. As David Kaye, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, noted in his 2016 report to the Human Rights Council, internet companies can affect rights when they engage in "overzealous censorship of a wide range of legitimate but (perhaps to some audiences) 'uncomfortable' expressions." David Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/32/38, ¶ 52 (May 11, 2016); *see also* Rikke Frank Jørgensen & Anja Møller Pedersen, *Online Service Providers as Human Rights Arbiters*, *in* THE RESPONSIBILITIES OF ONLINE SERVICE PROVIDERS 179 (Mariarosaria Taddeo & Luciano Floridi eds., 2017).

37. Molly K. Land, *Regulating Private Harms Online: Content Regulation Under Human Rights Law*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 285 (Rikke Frank Jørgensen ed., 2019).

38. MUNA ABBAS ET AL., INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA 7 (2019).

39. Charlie Campbell, *How China Is Using "Social Credit Scores" to Reward and Punish Its Citizens*, TIME (Jan. 16, 2019), https://time.com/collection/davos-2019/5502592/china-social-credit-score/.

40. Ironically, although this delegation is designed to further state interests, it often reduces state power in the long term by shifting authority to platforms. As David Kaye notes, "the pressure on companies has led to an outsourcing of public roles to private actors, which amounts to an expansion of corporate power instead of constraints on it." DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 20 (2019).

41. James Boyle, *A Nondelegation Doctrine for the Digital Age*, 50 DUKE L.J. 5, 10 (2000).

characterized the global web in order to carry out their own policy preferences."[42] The accountability gaps that result are significant:

> States are increasingly coercing online platforms and intermediaries to instantiate and enforce *public* policy preferences regarding online speech and privacy through *private* regulation—including not only ToS but also hash-sharing and other purportedly cooperative arrangements—that lacks critical accountability mechanisms. These coercive measures convert what might otherwise be private action into heterodox, hybrid public-private governance arrangements in which state and private power are commingled. In short, governments can avoid responsibility for their policy preferences if they force platforms to carry their water.[43]

States rely on intermediaries to regulate content for very practical reasons.[44] The sheer volume of content available online, combined with the challenge of identifying the source of such content and the difficulty of pursuing the actual violators across borders, makes policing online speech through traditional means costly and cumbersome.[45] Intermediaries, in contrast, are often easily controlled by states because they are engaged in business operations for which licensure may be required, or they may have physical assets or employees located in the country in question.[46]

Over the last four years, a series of events has led to increased concern about harms of speech on social media platforms and prompted further resort to platforms as proxy regulators. These events include terror attacks in the Europe and the United States, the Cambridge Analytica scandal, hate speech against refugees in the context of the Syrian refugee crisis, revelations in 2018 regarding the role of Facebook in online incitement to genocide against the Rohingya in Myanmar, and most recently the 2019 massacre at mosques in Christchurch, New Zealand.

Terrorist Attacks and Online Extremism: Terror attacks at the offices of Charlie Hebdo (January 2015), at the U.S. Navy Reserve in Chattanooga, Tennessee (July 2015), at several locations around the city of Paris

42. Bloch-Wehba, *supra* note 2, at 29.

43. *Id.* at 30.

44. Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 27-29 (2006). *Cf.* Jody Freeman, *The Private Role in Public Governance*, 75 N.Y.U. L. REV. 543, 580-81 (2000) (discussing the role and effects of private engagement in public governance).

45. MUELLER, *supra* note 20, at 149. Similar challenges were the impetus for notice and takedown liability in the United States as well as the creation of the ICANN domain name dispute process for trademark violations. *Id.* at 139-40.

46. RIKKE FRANK JØRGENSEN ET AL., CASE STUDY ON ICT AND HUMAN RIGHTS 4 (2015).

(November 2015), at the health department in San Bernardino, California (December 2015), and in the London Underground (December 2015) vaulted the issue of online extremism and its role in terrorist radicalization to the top of public consciousness. Subsequent attacks in 2016 and 2017 in Germany, Nice, Barcelona, London, Manchester, Paris, and Stockholm—among others—further heightened government pressure on intermediaries to remove extremist content feared to be contributing to radicalization.[47] The UK government in particular has been highly focused on the harms of online extremism, and in 2019 released the Online Harms White Paper, which details a set of recommendations for regulating online speech, including the creation of a new legal duty of care for social media platforms.[48]

Cambridge Analytica and Fake News: The Cambridge Analytica scandal in 2015-2016 led to heightened concern about the impact of misinformation and disinformation online. Starting in early 2015, journalists began reporting on how a UK data firm, Cambridge Analytica, appeared to be using private Facebook data to assist U.S. political campaigns.[49] The story gained international prominence, however, when details about the scope of the data harvesting—including the fact that it involved more than 50 million U.S. data profiles—were made public by a whistleblower.[50] Subsequent investigations have revealed that so-called "fake news" generated to target individuals on the basis of these profiles may have played a role not only in the UK referendum on leaving the European Union (Brexit) but also in bolstering support for the election of Donald Trump in the 2016 U.S. presidential election.[51] Concerns about the impact

47. Shirin Jaafari, *British Parliament Wants to Shut Down Extremist Content Online — At What Cost?*, PUBLIC RADIO INT'L: THE WORLD (Dec. 14, 2018), https://www.pri.org/stories/2018-12-14/british-parliament-wants-shut-down-extremist-content-online-what-cost; Molly Land, *The UK's Plan to Deny Terrorists 'Safe Spaces' Online Would Make Us All Less Safe in the Long Run*, CONVERSATION (June 15, 2017), https://theconversation.com/the-uks-plan-to-deny-terrorists-safe-spaces-online-would-make-us-all-less-safe-in-the-long-run-79323.

48. UK HOME OFFICE, ONLINE HARMS WHITE PAPER (2019), https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/793360/Online_Harms_White_Paper.pdf [hereinafter ONLINE HARMS WHITE PAPER].

49. Harry Davies, *Ted Cruz Using Firm that Harvested Data on Millions of Unwitting Facebook Users*, GUARDIAN (Dec. 11, 2015), https://www.theguardian.com/us-news/2015/dec/11/senator-ted-cruz-president-campaign-facebook-user-data.

50. Carole Cadwalladr & Emma Graham-Harrison, *"I Made Steve Bannon's Psychological Warfare Tool": Meet the Data War Whistleblower*, GUARDIAN (Mar. 18, 2018), https://www.theguardian.com/news/2018/mar/17/data-war-whistleblower-christopher-wylie-faceook-nix-bannon-trump.

51. Natasha Lomas, *Former Cambridge Analytica Director, Brittany Kaiser, Dumps More Evidence of Brexit's Democratic Trainwreck*, TECHCRUNCH (July 30, 2019), https://techcrunch.com/2019/07/30/brittany-kaiser-dumps-more-evidence-of-brexits-democratic-trainwreck/; Matthew Rosenberg et al., *How Trump Consultants Exploited the Facebook Data of Millions*, N.Y. TIMES (Mar. 17, 2018), https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html?module=inline.

of disinformation and misinformation have led governments to pressure social media companies to do more to remove so-called "fake news" and provide users with tools to verify online content.

Hate Speech and the Syrian Refugee Crisis: The rise of racist and xenophobic rhetoric online in connection with the influx of Syrian refugees fleeing to Europe in 2015-2016 and the potential connection of this hate speech to offline violence against refugees has led governments to call for social media companies to remove such content from their platforms. These concerns among others prompted Germany to pass a new law in 2017 that requires social media companies to remove unlawful content from their platforms.[52] Germany initially sought non-binding commitments from social media companies but moved to pass binding legislation after a German youth protection agency issued a report finding that the major companies only removed a portion of the illegal content that was flagged for them.[53] The new legislation, called the Netzdurchsetzunggesetz or NetzDG, imposes significant fines on companies with more than two million users in Germany that do not remove unlawful content within seven days of being notified of its presence on their platforms; this time frame is shortened to twenty-four hours for "manifestly unlawful" content.[54]

Incitement to Genocide in Myanmar: In late 2018, the UN-sponsored Independent International Fact-Finding Mission on Myanmar ("Fact-Finding Mission") released a report that concluded that online content shared on Facebook contributed both to specific acts of violence against the Rohingya minority in Myanmar as well as a general climate in Myanmar that made the genocide against the Rohingya possible.[55] By its own admission, Facebook was "too slow" to respond to the concerns of UN

---

52. Daniel Leisegang, *No Freedom to Hate: Germany's New Law Against Online Incitement*, EUROZINE (Sept. 29, 2017), https://www.eurozine.com/no-freedom-to-hate-germanys-new-law-on-online-incitement/; Morgan Meaker, *Can Governments Wrestle Power Back from Big Tech? Worldwide, Regulators Are Dealing with 'The Battle of Our Time,'* MEDIUM (Jan. 3, 2019), https://medium.com/s/story/can-governments-wrestle-power-back-from-big-tech-ae39e04334f3. A paper by Müller and Schwarz links anti-refugee content online to violence against refugees in Germany. *See* Karsten Müller & Carlo Schwarz, *Fanning the Flames of Hate: Social Media and Hate Crimes* (2019), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3082972.

53. WILLIAM ECHIKSON & OLIVIA KNODT, GERMANY'S NETZDG: A KEY TEST FOR COMBATTING ONLINE HATE 4-5 (2018); Leisegang, *supra* note 52.

54. Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [NetzDG] [Act to Improve Enforcement of the Law in Social Networks], Oct. 1, 2017, § 3(2)-(3) (Ger.), https://www.bmjv.de/SharedDocs/Gesetzgebungsverfahren/Dokumente/NetzDG_engl.pdf.

55. Indep. Int'l Fact-Finding Mission on Myan., Rep. of the Detailed Findings of the Independent International Fact-Finding Mission on Myanmar, U.N. Doc. A/HRC/39/CRP.2, ¶¶ 1310-1327 (Sept. 17, 2018) [hereinafter Myanmar Report]; *see also* Paul Mozur, *A Genocide Incited on Facebook, With Posts from Myanmar's Military*, N.Y. TIMES (Oct. 15, 2018), https://perma.cc/3Z2J-K6BA; Molly K. Land & Rebecca Hamilton, *Beyond Takedown: Expanding the Tool Kit for Responding to Online Hate*, *in* PROPAGANDA AND INTERNATIONAL CRIMINAL LAW: FROM COGNITION TO CRIMINALITY 143 (Predrag Dojčinović, ed., 2020).

officials and human rights advocates.[56] Facebook did not begin implementing a more robust and coordinated response to the problem until long after significant human rights violations had already occurred, prompting renewed calls for greater regulation of online platforms.

Christchurch, Extremism, and Online Violence: The 2019 massacre at two mosques in Christchurch, New Zealand reinvigorated government pressure on social media companies to remove extremist content. On March 15, 2019, a white supremacist opened fire on worshipers in two mosques in Christchurch, killing more than fifty people.[57] Much of the public debate focused on the way in which the perpetrator was able to incorporate social media platforms into his attack to promote his goals. One journalist observed, "[t]he attack was teased on Twitter, announced on the online message board 8chan and broadcast live on Facebook."[58] In short, as another journalist explained, this was an attack "engineered for internet virality."[59] The attacker's livestream of the massacres was quickly copied and disseminated, and efforts by Facebook, Twitter, YouTube, and others to remove it were not fully effective.[60] In responding to the massacre and the way in which social media was used by the perpetrator, the government of New Zealand released what it called the "Christchurch Call," a call for governments and social media companies to commit separately and collectively to a set of actions designed to address terrorist and violent extremist content online.[61]

---

56. Letter from Myanmar Civil Society Organizations to Mark Zuckerberg, Chairman & CEO, Facebook, Inc. (Apr. 5, 2018), https://perma.cc/3ZJB-XL3Z; Myanmar Report, *supra* note 55, ¶ 1351.

57. Eleanor Ainge Roy et al., *Christchurch Attack: Suspect Had White-Supremacist Symbols on Weapons*, GUARDIAN (Mar. 15, 2019), https://www.theguardian.com/world/2019/mar/15/christchurch-shooting-new-zealand-suspect-white-supremacist-symbols-weapons.

58. 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?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand; *see also* David D. Kirkpatrick, *Massacre Suspect Traveled the World but Lived on the Internet*, N.Y. TIMES (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/world/asia/new-zealand-shooting-brenton-tarrant.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand; Lois Beckett, *A History of Recent Attacks Linked to White Supremacy*, GUARDIAN (Mar. 15, 2019), https://www.theguardian.com/world/2019/mar/16/a-history-of-recent-attacks-linked-to-white-supremacism.

59. Hanna Ingber, *The New Zealand Attack Posed New Challenges for Journalists. Here Are the Decisions the Times Made*, N.Y. TIMES (Mar. 19, 2019), https://www.nytimes.com/2019/03/19/reader-center/new-zealand-media-coverage.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand.

60. Cade Metz & Adam Satariano, *Facebook Restricts Live Streaming After New Zealand Shooting*, N.Y. TIMES (May 14, 2019), https://www.nytimes.com/2019/05/14/technology/facebook-live-violent-content.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand; John Herrman, *The Internet's Endless Appetite for Death Video*, N.Y. TIMES (Mar. 24, 2019), https://www.nytimes.com/2019/03/24/style/really-bad-stuff.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand.

61. *Christchurch Call to Eliminate Terrorist and Violent Extremist Content Online*, https://www.christchurchcall.com/call.html [hereinafter *Christchurch Call*]; *see also* Charlotte Graham-

Each of these crises stimulated both public outcry and government pressure on social media companies to "clean up" their platforms. The next section catalogs the formal and informal means states are using to exert this pressure.

C. Techniques of Privatization

The internet and its foundational protocols were created and implemented with substantial public-private cooperation, and authority over central features of internet control was subsequently vested in private entities.[62] Further, public and private actors both vie for control over online content—private actors through contract and terms of service, and public actors through a range of techniques from legislation to court orders to informal pressure.

Most recently, however, states have been shifting their techniques of control toward deputizing private platforms to control online content on their behalf. This move is troubling because it circumvents domestic and international constraints on state activity and also disables existing mechanisms of accountability. For example, states are using these private regulatory systems to take down speech that they would not be allowed to remove themselves under national or international standards.[63]

Although others have focused on categorizing generations of control that rely on the technological infrastructure of online exchange,[64] this Article examines a new generation of legal and regulatory strategies. Broadly, there are three different types of approaches that states are taking in deputizing private intermediaries to regulate online content: command and control, intermediary liability, and extra-legal influence.[65] These are ideal types, however, and any given government action may exhibit

---

McLay & Adam Satariano, *New Zealand Seeks Global Support for Tougher Measures on Online Violence*, N.Y. TIMES (May 12, 2019), https://www.nytimes.com/2019/05/12/technology/ardern-macron-social-media-extremism.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand.

62. *See* MUELLER, *supra* note 20, at 61; Victoria D. Baranetsky, *Social Media and the Internet: A Story of Privatization*, 35 PACE L. REV. 304, 320-23 (2014).

63. KAYE, *supra* note 40, at 79 ("Governments could not order the removal of merely 'troubling' speech in the absence of a basis in national law, so they ask companies to do it instead.").

64. *See* Ronald Deibert & Rafal Rohozinski, *Beyond Denial: Introducing Next-Generation Information Access Controls*, *in* ACCESS CONTROLLED: THE SHAPING OF POWER, RIGHTS AND RULE IN CYBERSPACE 3, 4-7 (Ronald Deibert et al. eds., 2010) (first- and second-generation internet controls); Ronald Deibert, *Authoritarianism Goes Global: Cyberspace Under Siege*, 26 J. DEMOCRACY 64, 68-71 (2015) (third- and fourth-generation controls).

65. Keller categorizes state influence as consisting of regulation, indirect pressure, and cross-border influence. Keller, *supra* note 4, at 3. The categories used in this Article break down regulation and indirect pressure into their component parts to better analyze them under principles of state responsibility.

characteristics of more than one of these categories. Governments also use these techniques in combination with one another.

1. Command and Control

Some states control internet intermediaries directly. For example, some operators are directly state owned or controlled. Others may be independently owned but deeply intertwined with the state. For example, although the precise contours of China's relationships with its three largest internet intermediaries—Baidu, Alibaba, and Tencent—are unclear,[66] these companies are nonetheless "careful to demonstrate loyalty to the party" and "the government trusts them to heed [its] call to do whatever the regulatory bodies want."[67] In other instances, state agencies are directly involved in managing online content.[68]

In still other instances, governments do not control the activities of the intermediary directly but use the authority of the state to order them to take particular actions. Such orders might include court injunctions to block illegal file sharing sites or directing platforms to remove particular apps.[69] Commands might also include enforcement actions seeking to apply existing laws to the internet. Pakistan, for example, has used blasphemy laws to restrict access to Facebook, while in Lebanon, authorities used defamation law to justify the arrest of three Facebook users who had posted criticism of the Lebanese president.[70]

Command and control techniques are becoming increasingly sophisticated, as well. Some governments and international organizations are establishing law enforcement units that use platform terms of service enforcement to achieve their own ends. The Counter-Terrorism Internet Referral Unit (CTIRU) in the United Kingdom, for example, uses

66. For example, there were reports several years ago that the Chinese government planned to take stakes in tech companies that would allow it to "appoint government officials to the companies' boards and influence their operations," and that some companies had created "party committees . . . to make sure that firms do not stray from the path of socialism with Chinese characteristics." Masha Borak, *WSJ: Chinese Government Wants to Enter Boards of Chinese Tech Giants*, TECHNODE (Oct. 13, 2017), https://technode.com/2017/10/13/wsj-chinese-government-wants-to-enter-boards-of-chinese-tech-giants/; *see also* Li Yuan, *Beijing Pushes for a Direct Hand in China's Big Tech Firms*, WALL ST. J. (Oct. 11, 2017), https://www.wsj.com/articles/beijing-pushes-for-a-direct-hand-in-chinas-big-tech-firms-1507758314.

67. Emily Feng, *Chinese Tech Giants Like Baidu and Sina Set Up Communist Party Committees*, FIN. REV. (Oct. 11, 2017), https://www.afr.com/news/world/asia/chinese-tech-giants-like-baidu-and-sina-set-up-communist-party-committees-20171011-gyyh5u.

68. Gary King et al., *How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, Not Engaged Argument*, 111 AM. POL. SCI. REV. 484 (2017).

69. *See, e.g.*, Farhad Manjoo, *Clearing Out the App Stores: Government Censorship Made Easier*, N.Y. TIMES (Jan. 18, 2017), https://www.nytimes.com/2017/01/18/technology/clearing-out-the-app-stores-government-censorship-made-easier.html.

70. Ronald Deibert & Rafal Rohozinski, *Liberation vs. Control: The Future of Cyberspace*, 21 J. DEMOCRACY 43, 50-51 (2010).

providers' terms of service to initiate removal of content the government wishes to restrict. The CTIRU employs state agents to review content online; upon finding "extremist" content, agents notify the platform on which it was found that this content exists and constitutes a potential terms of service violation.[71] Europol launched an EU version of the CTIRU in July 2015,[72] and internet referral units (IRUs) have also been formed in the Netherlands, Belgium, Italy, and Germany.[73] The Israeli Cyber Unit similarly "appeals to content intermediaries like Facebook and Google to remove, restrict or suspend access to certain content, pages or users. These requests are based on an alleged violation of domestic laws as well as the intermediaries' own Terms of Service (ToS)."[74] IRUs vary in terms of the formality of their procedures and in particular whether the referring unit must first make a formal determination of illegality under national law before referring content to a platform.[75]

States are combining these referrals with legal penalties for non-compliance. The EU's proposed Regulation on terrorist content, for example, would require Member States to create a mechanism for referrals from law enforcement. Although the referrals themselves would not be legally binding, failure to comply with them exposes the platform to risk of sanction under the Regulation.[76] The UK White Paper, as well, provides that platforms must "[e]nforce their own relevant terms and conditions effectively and consistently"[77] to comply with their duty of care. A platform's failure to remove content referred to it by law enforcement as a violation of the platform's terms of service would likely constitute a breach of this duty of care.

2. Intermediary Liability

In this Article, intermediary liability refers to the liability of an intermediary for content appearing on its platform. By imposing liability, governments can regulate content indirectly by giving platforms an

71. Kaye, *supra* note 36, ¶ 53.

72. EUROPOL, EU INTERNET REFERRAL UNIT: TRANSPARENCY REPORT 2018, at 3 (2018).

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

74. *Israel State Attorney Claims Censorship of Social Media Content, Following Cyber Unit Requests, Isn't an 'Exercise of Gov't Authority*,' ADALAH (Nov. 28, 2019), https://www.adalah.org/en/content/view/9859.

75. Pielemeier & Sheehy, *supra* note 73.

76. Daphne Keller, *The EU's Terrorist Content Regulation: Expanding the Rule of Platform Terms of Service and Exporting Expression Restrictions from the EU's Most Conservative Member States*, CTR. FOR INTERNET & SOC'Y (Mar. 25, 2019), http://cyberlaw.stanford.edu/blog/2019/03/eus-terrorist-content-regulation-expanding-rule-platform-terms-service-and-exporting.

77. ONLINE HARMS WHITE PAPER, *supra* note 48, at 49.

incentive to act.[78] As Hiram A. Meléndez-Juarbe notes: "The objective of this strategy is to make these entities feel pressure in their pockets for potentially illegal activity of their clients and, in this way, use their technological resources to supervise or punish user activity (pursuant to the superior information and opportunity they have to do this, relative to governments or third parties)."[79] Liability can derive from explicit legislation or executive action, but can also arise as a result of failure to specify the liability of intermediaries, thus triggering (potentially uncertain) default background rules governing offline conduct.

Legal regimes that impose explicit liability on intermediaries generally have two principal elements.[80] First, they establish a standard of care (e.g., strict liability, recklessness, or negligence) by which the intermediary's actions will be evaluated. Second, they specify whether the intermediary has to proactively monitor content to avoid liability or whether it can wait until it is notified of potentially offending content before it must take action.

	Proactive Regime (Obligation to Monitor)	**Reactive Regime (Obligation after Notice)**
Strict Liability	• Chinese Cybersecurity Law • Thai Computer Crimes Act • European Court of Human Rights (clearly unlawful content)	• DMCA • NetzDG • EU e-Commerce Directive • Kenya's Computer Misuse and Cybercrimes Act • India's Information Technology Act & Copyright Act • Canada's "notice and notice" framework
Duty of Care	• UK White Paper (child	• UK White Paper (for all

78. MANILA PRINCIPLES ON INTERMEDIARY LIABILITY BACKGROUNDER PAPER 8 (2015), https://www.eff.org/sites/default/files/manila-principles-background-paper-0.99.pdf [hereinafter MANILA BACKGROUNDER].

79. Hiram A. Meléndez-Juarbe, *Intermediarios y libertad de expresión: apuntes para una conversación*, *in* HACIA UNA INTERNET LIBRE DE CENSURA: PROPUESTAS PARA AMÉRICA LATINA 109, 110 (Eduardo Bertoni ed., 2012) (English version at http://www.palermo.edu/cele/pdf/english/Internet-Free-of-Censorship/04-Intermediaries_Freedom_of_Expression_Hiram_Melendez_Juarbe.pdf).

80. Most who have surveyed this area have categorized intermediary liability into three types: strict liability, conditional liability, and broad immunity. *See, e.g.*, FOSTERING FREEDOM ONLINE, *supra* note 31, at 40-43. However, strict liability and conditional liability actually use the same standard of care (strict liability) but differ only in whether there is a safe harbor for intermediaries to avoid the imposition of this liability.

(duty to take reasonable steps)	pornography and extremist content)	other content)
Duty of Care (recklessness)	• Australian Abhorrent Violent Material Act	
Immunity	• Section 230 of the Communications Decency Act	

Strict Liability + Monitoring Obligation: In some jurisdictions, internet intermediaries are strictly liable for the content that appears on their platforms and must proactively police speech. The Chinese Cybersecurity Law of 2016, for example, prohibits dissemination of "false" information as well as information that disrupts national unity or national security and "requires companies to monitor their networks and report violations to the authorities"; failure to comply with the law can lead to heavy fines.[81] Intermediaries are liable "at every layer of a communication, from the ISP to the online service provider, website, and hosting company."[82] A company may face fines, criminal penalties, and loss of its license to do business if it "publishes or distributes content that regulators deem unlawful, or fails to sufficiently monitor the use of its services, take down content, or report violations."[83] Thailand's Computer Crimes Act similarly imposes strict liability plus a monitoring obligation for content prohibited by the government.[84]

At least with respect to clearly unlawful content, the European Court of Human Rights also approved, as consistent with the European Convention on Human Rights (ECHR), a standard of strict liability plus a monitoring obligation. In *Delfi AS v. Estonia*,[85] Delfi, an Internet news portal, argued that Estonia's decision to hold it responsible for defamatory content posted by its readers in the comments section violated Delfi's rights to freedom of expression under Article 10 of the ECHR.[86] The court found no violation in Estonia's decision to require Delfi to not only remove illegal content after being notified of its presence but also to affirmatively monitor user comments. According to the Court, member

81. David Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35, ¶ 15 (Apr. 6, 2018).

82. SHIELDING THE MESSENGERS, *supra* note 14, at 13; *see also* CYNTHIA WONG & JAMES X. DEMPSEY, MAPPING DIGITAL MEDIA: THE MEDIA AND LIABILITY FOR CONTENT ON THE INTERNET 17-18 (2011).

83. SHIELDING THE MESSENGERS, *supra* note 14, at 13.

84. CHARLES BRADLEY & RICHARD WINGFIELD, A RIGHTS-RESPECTING MODEL OF ONLINE CONTENT REGULATION BY PLATFORMS 31 (2018); *see also* WONG & DEMPSEY, *supra* note 82, at 18.

85. *See generally* Lisl Brunner, *The Liability of an Online Intermediary for Third Party Content: The Watchdog Becomes the Monitor: Intermediary Liability after* Delfi v Estonia, 16 HUM. RTS. L. REV. 1 (2016).

86. Delfi AS v. Estonia, 2015-II Eur. Ct. H.R. 319, ¶ 3.

states may "impose liability on Internet news portals, without contravening Article 10 of the Convention, if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties."[87]

Strict Liability + Safe Harbor: The other common approach employed to regulate intermediaries has been to create a regime of strict liability but to pair it with a safe harbor that allows the intermediary to avoid liability if it takes a particular action. A common form of conditional immunity is "notice and takedown," which has been used most frequently thus far to protect intellectual property rights online. In such regimes, intermediaries are liable for copyright violations only if they fail to remove the offending content after being notified of its presence on their systems. For example, in the United States, intermediaries are protected from liability for copyright violations if they remove offending content after they receive notice.[88] Under the European Union's e-Commerce Directive, "platforms are protected from legal liability for any illegal content they 'host' (rather than create) until they have either actual knowledge of it or are aware of facts or circumstances from which it would have been apparent that it was unlawful, and have failed to act 'expeditiously' to remove or disable access to it."[89] Canada uses a "notice and notice" system, in which notification of the presence of unlawful content only triggers an obligation to provide notice to the alleged offender, rather than an obligation to remove the content.[90]

Notice and takedown regimes are increasingly being used to regulate other harmful speech such as hate speech, fake news, and terroristic content. The new German network enforcement law (NetzDG), for example, requires social media companies to remove unlawful content after being notified of its presence on their networks or risk up to fifty million Euro in fines.[91] Section 56 of Kenya's 2018 Computer Misuse and Cybercrimes Act provides "that intermediaries will not be held responsible for any unlawful conduct unless it can be shown they had actual notice or actual knowledge of the conduct or that they acted wilfully and with malicious intent to facilitate the unlawful conduct."[92] In Vietnam, platforms must remove content within 24 hours of receiving notice from

87. *Id.* ¶ 159. A later case before the Fourth Section of the European Court of Human Rights indicates that this holding is likely limited to clearly unlawful content and then perhaps only in a commercial context. *See infra* notes 273-272 and accompanying text.

88. 17 U.S.C. § 512(c)(1)(C) (2018).

89. ONLINE HARMS WHITE PAPER, *supra* note 48, at 62.

90. SHIELDING THE MESSENGERS, *supra* note 14, at 20; *see also* OECD, *supra* note 11, at 6; MANILA BACKGROUNDER, *supra* note 78, at 17; MUELLER, *supra* note 20, at 138-39.

91. Kaye, *supra* note 81, ¶ 16.

92. John Walubengo & Mercy Mutemi, *Treatment of Kenya's Internet Intermediaries under the Computer Misuse and Cybercrimes Act, 2018*, 21 AFR. J. INFO. & COMMC'N 1 (2018).

the government.[93] India also follows the model of notice and takedown both with respect to copyright violations as well as in its Information Technology Act of 2000.[94]

Notice and takedown systems have long been under pressure, with many clamoring for governments to shift to proactive monitoring obligations. Copyright owners, for example, have put pressure on governments to limit the systems of conditional liability used for intellectual property violations, in order to require intermediaries to proactively monitor for infringement.[95] Recent legislative efforts in the United States have focused on making sites liable for turning a "blind eye" to obvious copyright infringement.[96]

Duty of Care + Monitoring Obligation: More recent efforts to impose liability on internet intermediaries combine a legal duty of care with an obligation to monitor. The UK White Paper, for example, proposes "a new statutory duty of care to make companies take more responsibility for the safety of their users and tackle harm caused by content or activity on their services."[97] This duty of care "will require companies to take reasonable steps to keep users safe, and prevent other persons coming to harm as a direct consequence of activity on their services."[98] The duty of care applies to both "illegal" activity as well as "harmful" (but not illegal) content.[99] With respect to extremist content and child pornography, the intermediary is responsible for content that it is not aware of and must proactively monitor its system to identify and remove this type of content.

The European Union is also moving toward a duty of care plus a monitoring obligation for extremist content. A 2018 proposal for a Regulation published by the European Commission and approved by EU Member States would establish a "minimum set of duties of care on hosting service providers."[100] The proposed Regulation would establish a requirement for EU Member States to create the legal infrastructure

93. Bethany Allen-Ebrahaimian, *Lawmakers Press Facebook and Google on Censoring U.S. Citizens for Vietnamese Government*, DAILY BEAST (July 17, 2018), https://www.thedailybeast.com/lawmakers-press-facebook-and-google-on-censoring-us-citizens-for-vietnamese-government.

94. SOFTWARE FREEDOM LAW CTR., INTERMEDIARY LIABILITY 2.0: A SHIFTING PARADIGM 3 (2019); Chinmayi Arun, *Gatekeeper Liability and Article 19(1)(A) of the Constitution Of India*, 7 NUJS L. REV. 73, 83 (2014).

95. Balkin, *supra* note 3, at 2321 (discussing efforts by the content industry to obtain legislation that would create liability for foreign sites that facilitate copyright infringement).

96. *Id.*

97. ONLINE HARMS WHITE PAPER, *supra* note 48, at 41; *see also* Nancy S. Kim, *Web Proprietorship and Online Harassment*, 2009 UTAH L.R. 993 (2009) (proposing amendments to Section 230 for foreseeable harm).

98. ONLINE HARMS WHITE PAPER, *supra* note 48, at 42.

99. *Id.*

100. *Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*, at 2, COM (2018) 640 final, 2018/0331 (COD) (Sept. 12, 2018) [hereinafter Proposed EU Regulation].

needed for a "removal order" which would be "issued as an administrative or judicial decision by a competent authority in a Member State."[101] When such an order is issued, "the hosting service provider is obliged to remove the content or disable access to it within one hour."[102] The proposed Regulation would also require "hosting service providers, where appropriate, to take proactive measures proportionate to the level of risk and to remove terrorist material from their services, including by deploying automated detection tools."[103] Thus, the proposed Regulation would, like the UK proposal, establish a quasi-negligence standard for service providers with respect to terrorist content combined with a proactive monitoring obligation.

Australia's new law, passed after the Christchurch massacre in New Zealand, imposes a duty of care but uses a recklessness standard, and it does not condition the duty on receipt of notice. The law makes it a criminal offense if a content service provider, whose platform can be used to access "abhorrent violent material," "does not ensure the expeditious removal of the material from the content service."[104] Although the Australian eSafety Commissioner can issue a notice designating content as "abhorrent violent material," the requirement to remove such material is not conditioned on notice. Receipt of such notice, however, will give rise to a presumption of recklessness, which is the mental state required for prosecution under the Act.[105]

Duty of Care + Safe Harbor: Although intermediaries would be required to proactively monitor for child pornography and terroristic content under the UK White Paper, they would only be responsible for other types of content after notice. According to the White Paper, "[t]he regulator will not compel companies to undertake general monitoring of all communications on their online services, as this would be a disproportionate burden on companies and would raise concerns about user privacy."[106] The paper defines notice to include identification by the intermediary itself of the allegedly offending content via technological means.[107]

Immunity: Governments also influence the conduct of intermediaries by granting them immunity from liability. As noted above, Section 230 of the

101. *Id.* at 4.

102. *Id.*

103. *Id.*

104. *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019*, s 474.34(1)-(4) (Austl.). *See generally* Evelyn Douek, *Australia's New Social Media Law Is a Mess*, LAWFARE (Apr. 10, 2019), https://www.lawfareblog.com/australias-new-social-media-law-mess.

105. *Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019*, s 474.35(5)-(6) (Austl.).

106. ONLINE HARMS WHITE PAPER, *supra* note 48, at 43.

107. *Id.* at 62.

Communications Decency Act was enacted to provide platforms with immunity in order to encourage intermediaries to voluntarily act to manage the content that appears on their sites.[108] Increased attention to the harms of online speech may be eroding support for the broad immunity conferred by Section 230.[109]

3. Extra-Legal Influence

States are also using a range of extra-legal techniques to pressure intermediaries to take particular actions with respect to content on their platforms. Although these techniques are embedded in law to varying degrees, I describe them here as "extra-legal" because they exist outside of normal legal process. There are three general ways in which this influence is currently being deployed: 1) systems of "voluntary self-regulation" developed under threat of government regulation; 2) partnerships and other institutional arrangements by which governments influence and even direct the actions of intermediaries; and 3) informal pressure, such as condemnation.

First, there are a number of systems of what is being called "voluntary self-regulation" that have been adopted by companies under the threat of state regulation. In Europe, for example, states concerned with harmful content online pressured companies to commit to a "Code of Conduct." This Code was negotiated in the fall of 2018, and it provides that Facebook, Microsoft, Twitter, and YouTube "have agreed with the European Commission on a code of conduct setting the following public commitments," which include, among others, the commitment to review the majority of notifications of illegal hate speech on their platforms and to remove or disable access to that content within twenty-four hours.[110] The Code of Conduct also requires the companies to provide EU Member States with information about how to submit notices of illegal conduct "with a view to improving the speed and effectiveness of communication between the Member State authorities and the IT Companies, in particular on notifications and on disabling access to or removal of illegal hate

108. 47 U.S.C. § 230(c)(2)(A)-(B) (2012).

109. For example, a bill proposed in March 2020, the Eliminating Abusive and Rampant Neglect of Interactive Technologies Act" or "EARN IT Act" would condition Section 230 immunity on compliance with particular "best practices" regarding user content management. Sophia Cope et al., *The EARN-IT Act Violates the Constitution*, Electronic Frontier Foundation (Mar. 31, 2020), https://www.eff.org/deeplinks/2020/03/earn-it-act-violates-constitution. *See also, e.g.*, Jeff Kosseff, *The Gradual Erosion of the Law that Shaped the Internet: Section 230's Evolution Over Two Decades*, 18 COLUM. SCI. & TECH. L. REV. 1 (2016).

110. *The EU Code of Conduct on Countering Illegal Hate Speech Online*, EUR. COMM'N, https://ec.europa.eu/info/policies/justice-and-fundamental-rights/combatting-discrimination/racism-and-xenophobia/countering-illegal-hate-speech-online_en.

speech online."[111] The agreement envisions that information about illegal content will be channeled through "national contact points designated by the IT companies and the Member States respectively."[112]

Second, there are a range of institutional (often explicitly governmental) arrangements by which governments seek to influence the conduct of intermediaries. The UK Council for Internet Safety, for example, is described on the UK government's web page as a "collaborative forum through which government, the tech community and the third sector work together to ensure the UK is the safest place in the world to be online."[113] The Better Internet for Kids Coalition was founded in 2011 by the European Union as a cooperative voluntary intervention aimed at making a "better and safer internet for children."[114] Among other things, it supports voluntary regulatory efforts, including the Safer Social Networking Principles, a self-regulatory agreement signed by social networking companies that do business in Europe, and the European Framework for Safer Mobile Use by Younger Teenagers and Children.

In the wake of the Christchurch massacre, the New Zealand government announced the "Christchurch Call," which is a call for governments and social media companies to commit separately and collectively to a set of actions designed to address terrorist and violent extremist content online.[115] The Christchurch Summit, which took place in May 2019, was a meeting organized by the New Zealand government and co-hosted by the French government designed to convene governments and technology companies "in an attempt to bring to an end the ability to use social media to organise and promote terrorism and violent extremism."[116]

Third, governments also regularly exert informal extra-legal pressure on internet intermediaries to take particular actions.[117] For example, in the

111. *Id.*

112. *Id.*

113. *What the UK Council for Internet Safety Does*, UK COUNCIL FOR INTERNET SAFETY, https://www.gov.uk/government/organisations/uk-council-for-internet-safety.

114. European Commission Press Release IP/11/1485, Digital Agenda: Coalition of Top Tech & Media Companies to Make Internet Better Place for Our Kids (Dec. 1, 2011), https://ec.europa.eu/commission/presscorner/detail/en/IP_11_1485.

115. *Christchurch Call*, *supra* note 61; *see also* Charlotte Graham-McLay & Adam Satariano, *New Zealand Seeks Global Support for Tougher Measures on Online Violence*, N.Y. TIMES (May 12, 2019), https://www.nytimes.com/2019/05/12/technology/ardern-macron-social-media-extremism.html?rref=collection%2Fspotlightcollection%2Fchristchurch-attack-new-zealand.

116. *The Christchurch Call: Helping Important Voices Be Heard*, INTERNETNZ, https://internetnz.nz/Christchurch-Call.

117. As Karanicolas notes, informal pressure or "jawboning" can be particularly problematic when it allows the state to circumvent constitutional limits. Michael Karanicolas, *Squaring the Circle Between Freedom of Expression and Platform Law*, XX J. TECH. L. & POL'Y 1, 11 (2019-2020). Bambauer describes state "persuasion" as a form of "soft" censorship. Bambauer, *supra* note 8, at 867. Deibert and Rohozinski call this category "informal requests" but also include slowdowns by state-owned

wake of controversy following the publication of the video "The Innocence of Muslims," many governments asked Google to remove the video.[118] Pressure from the Turkish government appears to have led to removals of pro-Kurdish content on Facebook.[119] Governments also pressure companies through public condemnation.[120] Following the WikiLeaks revelations, for example, the U.S. government pressured PayPal, Amazon, and others to cease doing business with WikiLeaks, and State Department Legal Advisor Harold Koh drafted a public letter that insinuated WikiLeaks had broken the law.[121] Companies quickly severed their business relationships with WikiLeaks, and Apple removed from its store an app that provided access to WikiLeaks documents.[122]

Such informal pressure on intermediaries is often combined with threats to impose greater legal regulation or liability. In the United States, intermediaries have been threatened with legislation and limits on their immunity to pressure them to voluntarily restrict access of repeat copyright infringers and adopt mandatory data retention laws.[123] The threat of liability under notice and takedown itself has also incentivized the creation of private partnerships for copyright enforcement.[124] Governments also use existing leverage points to obtain further concessions from intermediaries. In his 2018 report to the Human Rights Council, for example, David Kaye described the way in which the government in Pakistan used a three-year ban on YouTube to compel Google to "establish a local version susceptible to government demands for removals of 'offensive' content."[125] Indeed, the UK White Paper itself—before it is even enacted into law—might be seen as a form of government pressure backed up by the threat of regulation, since it seeks to exert pressure on companies to address online harms ahead of implementation.[126]

---

internet service providers and pressure from government officials to remove content. Deibert & Rohozinski, *supra* note 70, at 51.

118. DENARDIS, *supra* note 6, at 158.

119. Sara Spary, *Facebook Is Embroiled in a Row With Activists Over "Censorship"*, BUZZFEED NEWS (Apr. 8, 2016), https://www.buzzfeed.com/saraspary/facebook-in-dispute-with-pro-kurdish-activists-over-deleted#.cj5EVrNLJ.

120. Bambauer, *supra* note 8, at 891-99; *see also, e.g.*, Robert Hutton, *U.K. Warns Facebook, Google, Twitter to Better Fight Hate Speech*, BLOOMBERG (Apr. 20, 2017), https://www.bloomberg.com/news/articles/2017-04-30/u-k-warns-facebook-google-twitter-to-better-fight-hate-speech.

121. DENARDIS, *supra* note 6, at 162; Balkin, *supra* note 3, at 2327-28.

122. Balkin, *supra* note 3, at 2328.

123. Bambauer, *supra* note 8, at 896-97.

124. Bridy, *Graduated Response*, *supra* note 9, at 85.

125. Kaye, *supra* note 81, ¶ 20.

126. ONLINE HARMS WHITE PAPER, *supra* note 48, at 27.

III. PRIVATIZED CENSORSHIP UNDER INTERNATIONAL LAW

States are increasingly turning to intermediaries not just as effective points of control, but as regulators. This delegation of authority to intermediaries, along with the loss of transparency and accountability the shift entails, is highly problematic under both domestic and international norms regarding rule of law.

This section considers the lawfulness of this move toward privatized speech governance. It first examines the state action doctrine under international law, which in most instances imposes duties only on public actors. It then considers how each of the new forms of state control discussed above fare under the state action rules of international law. Finally, this section also makes the case for the development of a more flexible state action doctrine derived from principles of state responsibility for purposes of determining when content moderation must comply with human rights law.

This Article relies on international law as a baseline for several reasons. First, international law is the law that applies to the conduct of governments, and the move toward privatized censorship is an agenda that is being pushed by states—which are the primary subjects of international law. Second, as David Kaye, the UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Expression and Opinion, has argued, international human rights is a reasonable, legitimate, and effective global baseline for evaluating the activities of global social media platforms.[127] It is not appropriate for global media platforms to privilege one national approach to speech over another.[128] Nor should platforms adopt their own rules untethered by common norms. Rather, as Kaye explains, human rights law "offer[s] a globally recognized framework for designing those tools and a common vocabulary for explaining their nature, purpose and application to users and States."[129]

A. Non-State Actors

For human rights law and institutions, private control of online content presents a conundrum. Both state and private actors can cause human rights harms online, but international human rights law generally

127. Kaye, *supra* note 81, ¶¶ 41-43; *see also* Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939, 966-68 (2020); Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 34-35 (2018).

128. Kate Klonick has documented the way in which U.S. First Amendment law influenced the work of Silicon Valley-based platforms. *See* Klonick, *supra* note 1, at 1621.

129. Kaye, *supra* note 81, ¶ 43.

only regulates the former. Under normative principles developed by then UN Special Representative on Business and Human Rights, John Ruggie, and endorsed by the United Nations Human Rights Council in the form of the *UN Guiding Principles on Business and Human Rights*, non-state actors have a moral—but not a legal—obligation to respect human rights in their activities.[130]

Human rights law typically responds to the problem of regulating the human rights impact of private companies through a combination of national regulation and nonbinding frameworks. First, human rights law imposes a legal obligation on the *state* to protect individuals from harm and also provide remedies when rights have been violated.[131] The *Guiding Principles* emphasize that "[s]tates must protect against human rights abuse within their territory and/or jurisdiction" and toward this goal must take "appropriate steps to prevent, investigate, punish and redress such abuse through effective policies, legislation, regulations and adjudication."[132] This positive obligation is also at the heart of the new draft treaty on business and human rights, which is currently being negotiated under the auspices of the United Nations. Article 10(1) of the draft provides: "State Parties shall ensure through their domestic law that natural and legal persons may be held criminally, civil or administratively liable for violations of human rights undertaken in the context of business activities of transnational character."[133]

Second, the Guiding Principles articulate a set of non-binding responsibilities for business. This corporate responsibility to respect rights includes both the duty to "[a]void causing or contributing to adverse human rights impacts through their activities" and to "[s]eek to prevent or mitigate adverse human rights impacts that are directly linked to their operations."[134] Toward these goals, the *Guiding Principles* state that companies must implement a policy commitment to meet their

130. UNITED NATIONS OFF. OF THE HIGH COMMISSIONER FOR HUM. RTS., GUIDING PRINCIPLES ON BUSINESS AND HUMAN RIGHTS: IMPLEMENTING THE UNITED NATIONS "PROTECT, RESPECT AND REMEDY" FRAMEWORK 13 (2011) [hereinafter GUIDING PRINCIPLES]; *see also* John Ruggie, *Protect, Respect and Remedy: A Framework for Business and Human Rights, Report of the Special Representative of the Secretary-General on the Issue of Human Rights and Transnational Corporations and Other Business Enterprises*, ¶ 64, U.N. Doc. A/HRC/8/5 (Apr. 7, 2008). *See generally* Philip Alston, *The "Not-a-Cat" Syndrome: Can the International Human Rights Regime Accommodate Non-State Actors?*, *in* NON-STATE ACTORS AND HUMAN RIGHTS 1, 36 (Philip Alston ed., 2005).

131. GUIDING PRINCIPLES, *supra* note 130, at 3, 27.

132. *Id.* at 3.

133. UNITED NATIONS OFF. OF THE HIGH COMMISSIONER FOR HUM. RTS., LEGALLY BINDING INSTRUMENT TO REGULATE, IN INTERNATIONAL HUMAN RIGHTS LAW, THE ACTIVITIES OF TRANSNATIONAL CORPORATIONS AND OTHER BUSINESS ENTERPRISES, ZERO DRAFT (2018), https://www.ohchr.org/Documents/HRBodies/HRCouncil/WGTransCorp/Session3/DraftLBI.pdf.

134. GUIDING PRINCIPLES, *supra* note 130, at 14.

responsibilities, establish a system for engaging in human rights due diligence in order to "identify, prevent, mitigate, and account" for their human rights impacts, and create processes to remediate harms.[135] The *Guiding Principles* offer additional detail about the nature of the policy commitment, due diligence, and remediation systems required,[136] and also emphasize that companies should comply with all applicable laws, honor international human rights standards, and treat human rights harms as an issue of legal compliance.[137] The *Guiding Principles* are non-binding, however, and have been critiqued as insufficiently robust to result in meaningful changes in corporate behavior.[138] Nonetheless, they represent at the very least a "widely accepted framework for managing the behaviors of business activities that may impact human rights."[139]

Increasingly, scholars and activists alike are discussing what corporate respect for human rights might look like in the context of internet governance. For example, in considering the issue of content moderation, is the decision of a private platform to remove a particular piece of content a violation of freedom of expression? Emily Laidlaw, for example, asks:

> When a platform deletes user content because it infringes the Terms of Service, how is this to be framed? Do users have a right to freedom of expression on a private platform, or does the company have a corresponding duty to respect user rights? Should platforms match the approach of governments in delineating limits to these rights, or should it carve out stricter rules on the basis of social responsibility?[140]

The state action requirement in human rights law—the fact that international law in general only binds states and not private actors—means as a formal matter that content removed by private actors does not trigger human rights scrutiny.[141] This would mean, then, that a platform

---

135. *Id.* at 15-16.

136. *Id.* at 16-25.

137. *Id.* at 25-26.

138. Larry Catá Backer, *Moving Forward the UN Guiding Principles for Business and Human Rights: Between Enterprise Social Norm, State Domestic Legal Orders, and the Treaty Law That Might Bind Them All*, 38 FORDHAM INT'L L.J. 457, 461-62 (2015).

139. *Id.* at 458.

140. Emily B. Laidlaw, *Online Platform Responsibility and Human Rights*, *in* PLATFORM REGULATIONS: HOW PLATFORMS ARE REGULATED AND HOW THEY REGULATE US 65, 66 (Luca Belli & Nicolo Zingales eds., 2017); *see also* Bloch-Wehba, *supra* note 2, at 56 ("In fact, although the European Court of Justice concluded in the Telekabel case that users have standing to contest over-removal that results from judicial action, there is no clear path to do so when the removal appears to result from private action.").

141. The formalist distinction in international law between public and private is similar in orientation to the U.S. state action doctrine, which is "premised on the classical conception of

would be entirely free to decide what content it does and does not want to disseminate, and it would not be obligated to provide any explanation for or justification of those decisions. From the perspective of users, however, the fact that expression is silenced by a private rather than a public actor does not matter. Apple's decision to remove from the iTunes App Store an application that distributed publicly available data about drone strikes,[142] or Facebook's deletion of a Syrian artist's photos of refugees who had drowned off the coast of Libya[143]—look and feel like censorship. A report from the project Onlinecensorship.org, which collects information from users about content removals, discusses the public interest impact of these removals. Among other things, content removed during the eight months studied in 2016 included a well-known Vietnam-era photo of a naked girl fleeing a napalm attack[144] and videos of police shootings (including Philando Castile and Korryn Gaines).[145] Thirty-two percent of the data studied related to the U.S. elections, and "[m]any of these users reported extreme frustration with the removal of their content, as they sought to speak their minds and share information about the highly contested election."[146] For individuals affected by such decisions, it is of little comfort that their content was blocked or removed by a private rather than a public actor. And when there are no other meaningful avenues for that expression, the private removal has the force of law.

Some have sought to respond to these problems using gatekeeper theory. Under gatekeeper theory, particular intermediaries may have special responsibilities by virtue of their dominance, status, or influence on democracy.[147] Laidlaw, for example, argues that companies with greater impacts on democracy should have greater obligations to ensure that discourse can take place. According to Laidlaw, the human rights responsibilities of a company "increase or decrease based on the extent

---

powers that are autonomous within their spheres." *Developments in the Law: State Action and the Public/Private Distinction*, 123 HARV. L. REV. 1248, 1256 (2010) [hereinafter *Developments in the Law*].

142. Sam Biddle, *Apple: Drone Strikes Are Offensive, Farts and Poop Are Cool*, GAWKER (Sept. 28, 2015), http://gawker.com/apple-kills-drone-strike-news-app-for-being-too-crude-1733402994.

143. Nicholas D. Mirzoeff, *Facebook Censors Refugee Photographs*, HOW TO SEE THE WORLD (Sept. 1, 2015), https://wp.nyu.edu/howtoseetheworld/2015/09/01/auto-draft-78/.

144. JESSICA ANDERSON ET AL., CENSORSHIP IN CONTEXT: INSIGHTS FROM CROWDSOURCED DATA ON SOCIAL MEDIA CENSORSHIP 10 (2016).

145. *Id.* at 13. Kyle Langvardt discusses the risks to freedom of expression posed by intermediaries acting as content regulators. Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1355-56 (2018).

146. ANDERSON ET AL., *supra* note 144, at 18-19.

147. EMILY B. LAIDLAW, REGULATING SPEECH IN CYBERSPACE, GATEKEEPERS, HUMAN RIGHTS AND CORPORATE RESPONSIBILITY 46 (2016) ("The human rights framework proposed here depends on the extent to which the gatekeeper controls deliberation and participation in the forms of meaning-making in democratic culture."); *see also* Jonathan Zittrain, *A History of Online Gatekeeping*, 19 HARV. J.L. & TECH. 253 (2006); Reinier H. Kraakman, *Gatekeepers: The Anatomy of a Third-Party Enforcement Strategy*, II:1 J. L., ECON. & ORG. 53 (1986).

that its activities facilitate or hinder democratic culture."[148] Jørgensen similarly asks whether Google should be treated as a private company or whether the importance of its services mean it has "an extra obligation to respect human rights standards."[149] Gatekeepers may be defined both by reference to market dominance as well as the extent to which there are reasonably adequate alternatives to the platform or services they provide.[150]

Gatekeeper theory makes intuitive sense and can be well justified by reference to moral, ethical, and political arguments. Arguments grounded in political theory for special responsibilities of gatekeepers can be paired with legal arguments based on the language of Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which imposes some direct legal duties on dominant intermediaries.[151] National law may also be a source of obligation. Relying on the *Drittwirkung* doctrine, German courts in several 2018 cases found that Facebook was required to "observe fundamental rights when it determines whether to delete content pursuant to its ToS."[152]

Even when they are directly bound by international or national law, however, it is unclear how that law should apply when regulating freedom of expression. Even the most dominant of intermediaries should be able to take into account their own business needs and the preferences of the users of their platforms, including in eliminating offensive content.[153] More work is needed to understand how the principles of human rights law should apply on these large platforms.[154]

Others have sought to move beyond formal distinctions of public and private. Hanna Bloch-Wehba, for example, argues that platforms should adopt basic principles of administrative law to ensure accountability to the public.[155] David Kaye, Barrie Sander, Evelyn Aswad, and Michael Karanicolas recommend that platforms adopt international human rights standards.[156] Kaye explains:

148. LAIDLAW, *supra* note 147, at 48.

149. RIKKE FRANK JØRGENSEN, FRAMING THE NET: THE INTERNET AND HUMAN RIGHTS 95 (2013).

150. Land, *supra* note 37, at 304.

151. *Id.* at 303-04; *see also* Molly Land, *Toward an International Law of the Internet*, 54 HARV. J. INT'L L. 393 (2013); *cf.* Langvardt, *supra* note 145, at 1371-72 (exploring a variety of ways in which domestic legislation might draw distinctions between intermediaries).

152. Bloch-Wehba, *supra* note 2, at 77.

153. Karanicolas, *supra* note 117, at 24 (discussing the right of platforms "not to speak").

154. *See generally* Kaye, *supra* note 36; Sander, *supra* note 127; Aswad, *supra* note 127; Karanicolas, *supra* note 117.

155. Bloch-Wehba, *supra* note 2, at 33.

156. Kaye, *supra* note 36, ¶¶ 41-43; Sander, *supra* note 127, at 20-22; Aswad, *supra* note 127, at 57-67; Karanicolas, *supra* note 117, at 25.

> Some argue that human rights law applies only to governments and not to companies. But that is rapidly becoming an archaic way of thinking about the structure of international governance. There is a growing recognition that corporations have responsibilities not to interfere with the rights individuals enjoy, whether it is a multinational company involved in mineral extraction that helps fuel conflict or undermine worker rights, or an internet company sharing user data with an authoritarian regime.[157]

These scholars advance persuasive policy arguments for why digital intermediaries should follow human rights law in their activities. And given the lack of coercive mechanisms for enforcing international law, the distinction between a legal obligation and a moral responsibility may have little practical impact.

Nonetheless, distinguishing between public and private authority is important in the context of content moderation for several reasons. Although it does seem clear that platforms have human rights responsibilities, it is not at all clear what those responsibilities are or when they are triggered. First, how do we identify a violation in the absence of state action? Although a government clearly could not suppress speech simply because it embodies a viewpoint on one end of the ideological spectrum or the other, it is not necessarily evident that private platforms are prohibited from picking and choosing in this way. In other words, it is difficult to determine when private actors violate the right to freedom of expression because state action is itself an element of the violation.[158]

Second, it is not evident when these responsibilities might be triggered. The nature of the obligations assumed would have to vary based on the size and dominance of the platform. When applied to an individual blog, for example, the cost of requiring compliance with norms of free expression, when weighed against the marginal harm to a user, would be

157. KAYE, *supra* note 40, at 119-20.

158. *Cf. Developments in the Law*, *supra* note 141, at 1255 ("The state action doctrine establishes a threshold requirement for judicial consideration of constitutional claims and congressional enforcement of constitutional rights: absent some action on the part of a state entity, the doctrine holds, there can be no constitutional violation."). Sander argues that this might be resolved by reliance on the positive obligation of the state to regulate. *See* Barrie Sander, *Democratic Disruption in the Digital Age: Social Media, Cyber Governance and International Human Rights Law* (forthcoming 2020) (manuscript at 8-9) ("Under IHRL, the private censorship practices of social media companies – whether influenced by informal governmental pressures or commercial interests – generally fall to be examined in terms of the positive obligations of States to ensure that persons within their territory and/or jurisdiction are protected from acts of private actors that would impair their enjoyment of the right to freedom of expression."). Identifying impairments to the enjoyment of the right to freedom of expression even for this more limited purpose, however, is difficult absent greater clarity about what constitutes a violation and when platform responsibility is triggered.

disproportionate.[159] This inquiry is complicated by the fact that public and private action are deeply intertwined in this space. It is not sufficient to look only at the identity of the actor that restricted the content, because some of what might otherwise appear as private action may be state action in disguise. Adapting rules of state responsibility to help untangle public and private authority in the online context could help us distinguish between what is truly private (although perhaps still subject to non-binding human rights responsibilities) and what should be treated as public action that *must* comply with human rights law.

Thus, principles of state responsibility are being used here not to transform content moderation decisions into state actions triggering international responsibility, but rather to distinguish between those actions of a platform that should be considered "private" and those which entail a sufficient level of public authority to require respect for human rights law. The goal, in this endeavor, is to prevent states from leveraging private actors to do their bidding while avoiding accountability.[160] As Jørgensen and Pedersen explain, "intermediaries are being used to implement public policy with limited oversight and accountability with severe implications on human rights."[161] Among other things, this allows states to "de facto neglect their human rights obligations and escape the strict requirements, which would have been otherwise incumbent upon them had they applied the restrictions themselves."[162] Thus, we must begin to disentangle the "public" and the "private" in order to ensure that states cannot so easily evade their existing obligations under human rights law.

Under human rights law, content removals must pursue a legitimate purpose, be established in law, and be necessary and proportional to the ends to be achieved.[163] This does not mean that governments cannot or should not rely on intermediaries in their efforts to respond to the harmful effects of online speech. Indeed, it is unlikely that this task could be accomplished without some measure of public-private cooperation. It does

159. This diversity of online providers is why the spatially-based approach of U.S. law, which applies the First Amendment to speech that occurs in "locations traditionally understood as public, even if they are not publicly owned," *Developments in the Law*, *supra* note 141, at 1303, will not work well in the online context.

160. Indeed, as Andrew Clapham notes, privatization is often deployed by the state precisely in order to insulate it from accountability. Andrew Clapham, *Non-State Actors*, *in* INTERNATIONAL HUMAN RIGHTS LAW 531, 534 (Daniel Moeckli et al. eds., 2d ed. 2006) ("First, the whole point of privatization is to remove certain activity from the state-run sphere, increasing flexibility and usually diminishing accountability. The usual consequence is that the protections that have developed with regard to state activity no longer apply and, although in theory new private remedies should apply, such remedies tend not to include human rights protection in name.").

161. Jørgensen & Pedersen, *supra* note 36, at 180.

162. *Id.*; *see also* JØRGENSEN, *supra* note 149, 98-100.

163. Frank La Rue, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 24, U.N. Doc. A/HRC/17/27 (May 16, 2011).

mean, however, that states cannot delegate public authority to private entities and thereby evade lawful human rights, constitutional, and rule-of-law checks.[164] States may not rely on gatekeepers to obscure and deflect attention from the policies they are seeking to advance,[165] or to engage in what Milton Mueller and Daphne Keller have called policy laundering.[166] It means states cannot require intermediaries to take down content that the state itself would not be allowed to take down, or to impose penalties that it would not be able to impose.

B. Principles of State Responsibility Online

Under international law, there are limited instances in which private action may be directly attributable to the state under principles of state responsibility. The International Law Commission's Articles on State Responsibility, which the UN General Assembly commended to the attention of states in 2001, provides a framework for understanding the conditions under which action by non-state entities can be attributable to the state for purposes of state responsibility. Broadly, there are two main provisions of the Articles on State Responsibility potentially relevant to understanding when the actions of a private internet intermediary must comply with human rights law. Article 8 describes circumstances in which a private entity acts under the "instruction, direction, or control" of a government, and Article 5 discusses situations in which private actors exercise governmental authority.[167]

Principles of state responsibility are not usually applied to establish responsibility for human rights violations. Rather, under human rights law, states are indirectly responsible for the actions of private actors when they systematically fail to protect individuals from harms those actors cause.[168] This form of liability is not derivative liability for the acts of a private actor but rather a new and independent wrongful act by the state—namely, the

---

164. Bambauer notes that "the government may push intermediaries to censor speech that it could not lawfully block itself," which would "insulate[] state efforts from constitutional challenge, since private parties formally make the decisions regarding content." Bambauer, *supra* note 8, at 897.

165. Balkin, *supra* note 3, at 2297-298, 2304.

166. MUELLER, *supra* note 20, at 211; Keller, *supra* note 4, at 3; *see also* SHIELDING THE MESSENGERS, *supra* note 14, at 4.

167. G.A. Res. 56/83, annex, Responsibilities of States for Internationally Wrongful Acts, arts. 8, 5 (Dec. 12, 2001) [hereinafter *Articles on State Responsibility*]. To the extent that the action in question is that of the European Union, an international organization, the relevant principles of attribution are reflected in the Draft Articles on the Responsibility of International Organizations, adopted by the International Law Commission and submitted to the General Assembly. Rep. of the Int'l Law Comm'n, Rep of the Work of its Sixty-Third Session, U.N. Doc. A/66/10, ¶ 87 (2011). Those rules differ from the ones that apply to states in ways that are beyond the scope of this analysis.

168. Velásquez Rodriguez v. Honduras, Merits, Reparations and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 4, ¶ 182 (July 21, 1989).

failure to prevent, prosecute, and punish harms to rights perpetrated by non-state actors.[169] Principles of state responsibility, in contrast, are used to determine when a state should be held responsible for a breach of an obligation to another state, including the conditions under which an action by a non-state actor should be attributed to the state for that purpose.

However, the deeply intertwined nature of public and private authority online requires rethinking the traditional approach to state responsibility in the context of human rights law. There is a fundamental mismatch between frameworks of human rights accountability that emphasize obligations to prevent and punish and the reality on the ground where public actors are leveraging private authority to achieve their goals. In light of these accountability challenges, scholars have begun to resist a formalist approach that rigidly distinguishes between public and private action, in favor of a more functionalist approach that can better respond to accountability gaps.[170] Kate Crawford and Jason Schultz, for example, have recently argued that traditional state action doctrines can be used to hold developers of AI systems accountable for constitutional violations caused by their systems.[171] Discussing the "right to be forgotten" in the European Union, Edward Lee argues that Google should be viewed as a "private administrative agency" because of the authority it has been delegated under EU law and the public functions it performs.[172]

International scholars, as well, have begun to make similar arguments about the need to update the state action doctrine to meet the demands of contemporary challenges. Antonio Cassese, for example, argues that a more flexible approach to principles of state responsibility is needed in light of deeply intertwined public and private action in the area of military activity, including state support of military and paramilitary groups,

---

169. Kubo Mačák, *Decoding Article 8 of the International Law Commission's Articles on State Responsibility: Attribution of Cyber Operations by Non-State Actors*, 21 J. CONFLICT & SEC. L. 405, 428 (2016); Jan Arno Hessbruegge, *The Historical Development of the Doctrines of Attribution and Due Diligence in International Law*, 36 N.Y.U J. INT'L L. & POL. 265, 268 (2004).

170. Gillian E. Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1373 (2003) ("The inadequacies of current state action doctrine mean that private exercises of government power are largely immune from constitutional scrutiny, and therefore expanding privatization poses a serious threat to the principle of constitutionally accountable government."); Kiel Brennan-Marquez, *The Constitutional Limits of Private Surveillance*, 66 KAN. L. REV. 485, 511 (2018) (private action may be considered state action when the government "capitaliz[es] on the activity of private companies" in order to "expand[] its infrastructural capability"); *see also* Rory Van Loo, *Rise of the Digital Regulator*, 66 DUKE L.J. 1267 (2017). Indeed, U.S. law has shifted in the past to ensure formalism does not inhibit important public policy goals. *Developments in the Law*, *supra* note 141, at 1259 (discussing the shift in U.S. state action doctrine as a result of the failure of formalist public/private distinctions to address racial discrimination).

171. Kate Crawford & Jason Schultz, *AI Systems as State Actors*, 119 COLUM. L. REV. 1941, 1943-44 (2019).

172. Edward Lee, *Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten*, 49 U.C. DAVIS L. REV. 1017, 1055, 1066, 1070 (2016).

increasing state support of terrorist groups, and the use of national military contingents by international organizations.[173] In the context of cybersecurity, which poses particular attribution challenges, Mačák argues that international law may be moving toward a more permissive interpretation of "control," either in general or on a case by case basis.[174] As Cassese emphasizes, the purpose of the rules of state responsibility is to ensure that

> states may not evade responsibility towards other states when they, instead of acting through their own officials, use groups of individuals to undertake actions that are intended to damage, or in the event do damage, other states; if states so behave, they must answer for the actions of those individuals, even if such individuals have gone beyond their mandate or agreed upon tasks—lest the worst abuses should go unchecked.[175]

In the context of online content regulation, public and private action is deeply enmeshed in ways that trigger precisely the policy concerns voiced by Cassese. For example, a 2015 study of intermediary self-regulation online argues that "too rigid an understanding of the 'state action'-doctrine would not do justice to the varying degrees of, and degrees of complexity of, state involvement in self-regulatory measures."[176] Although states would of course remain responsible for their own actions, rigid distinctions between public and private would mean that the actual exercise of the delegated authority by private entities engaged in online speech regulation would be insulated from accountability. In light of national variation in determining what constitutes governmental action,[177] recourse to an international standard could also provide more predictability.

Finally, a more flexible state action doctrine is also required in the context of freedom of expression because state action is a threshold issue for determining the existence of a violation. Aside from direct responsibilities that may operate by virtue of a platform's size or dominance in the market, a private platform that removes political speech

---

173. Antonio Cassese, *The Nicaragua and Tadić Tests Revisited in Light of the ICJ Judgment on Genocide in Bosnia*, 18 EUR. J. INT'L L. 649, 665-67 (2007).

174. Mačák, *supra* note 169, at 423-25.

175. Cassese, *supra* note 173, at 654; *see also* Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶ 117 (Int'l Crim. Trib. for the Former Yugoslavia July 15, 1999) (advancing support for different levels of rigor in the control test threshold).

176. C.J. Angelopoulos et al., *Study of Fundamental Rights Limitations for Online Enforcement Through Self-Regulation*, at 51 (Inst. for Info. Law 2015) [hereinafter *IViR Study*].

177. *See* ADALAH, *supra* note 74 (State Attorney's Office in Israel determined that requests originating from the Israeli Cyber Unit "do not constitute an exercise of governmental authority").

has not formally violated the speaker's human rights. Thus, in order to determine whether there has been a violation at all, it is necessary to understand whether the speech limitation in question is purely "private" or whether it was in fact done pursuant to public authority. Indeed, human rights tribunals have adopted a more lenient state action doctrine when needed to determine whether there has been a violation. In the Mapiripán Case, for example, the Inter-American Court of Human Rights found sufficient state action based on the fact that public authorities tolerated or supported the paramilitary groups that committed the violations in question.[178] This is a less stringent test for state action, but it raises fewer concerns in this context because it is a threshold determination decoupled from traditional enforcement mechanisms for state-to-state breaches of international obligations. Thus, a more flexible approach to state action could be used at the very least to identify when the exercise of delegated authority for speech regulation should be governed by international law, even if it does not give provide the basis for state-to-state remedies in the traditional sense.

C. Applying the State Action Doctrine

This section examines the ways in which states are using platforms to regulate speech and considers how established principles of international law would apply to these activities.[179] It uses the classifications set out in Part II—command and control, intermediary liability, and extra-legal influence—and analyzes each of these types of activities under principles of state responsibility.

1. Command and Control

Command and control activities involve instances in which governments either exert control over intermediaries directly or use the authority of the state to require intermediaries to take particular actions. It also includes the creation of internet referral units, such as the CTIRU in the United Kingdom and similar units throughout Europe, which identify

178. Anja Seibert-Fohr, *From Complicity to Due Diligence: When Do States Incur Responsibility for their Involvement in Serious International Wrongdoing?*, 60 GERMAN Y.B. INT'L L. 667, 679 (2017). This approach might be satisfied on a showing of what the U.S. Supreme Court has called "pervasive entwinement." Metzger, *supra* note 170, at 1415; *see also* Brentwood Acad. v. Tenn. Secondary Sch. Athletic Ass'n, 531 U.S. 288, 288 (2001).

179. Internet service providers are unlikely to be deemed state actors for purposes of the U.S. First Amendment. Jonathan Peters, *The "Sovereigns of Cyberspace" and State Action: The First Amendment's Application-Or Lack Thereof-To Third-Party Platforms*, 32 BERK. TECH. L.J. 989, 992 (2018); *see also* Enrique Armijo, *Government-Provided Internet Access: Terms of Service as Speech Rules*, 41 FORDHAM URB. L.J. 1499 (2015) (discussing the state action doctrine as it would apply to government-provided internet service).

problematic content and notify the platform that this content may violate its terms of service. Failure to comply with these referrals is sometimes backed up by sanctions.

Under Article 8 of the Articles on State Responsibility, states are responsible for wrongful acts done pursuant to their instruction, direction, or control.[180] This provision addresses situations in which private parties carry out wrongful conduct pursuant to instructions or directions issued by the state, or where the state exercises effective control over the private party.[181] Both tests require a fairly high showing to establish state responsibility.

The test for "control," for example, requires "effective control" over the wrongdoing itself.[182] This is commonly understood to be a stringent test by which the state must in essence "micromanage[] the conduct of the individual agents."[183] Although the International Criminal Tribunal for the Former Yugoslavia (ICTY) applied a lower standard in attributing responsibility in order to establish the existence of an international conflict,[184] the International Court of Justice rejected this standard for the purposes of general attribution in the *Bosnian Genocide* case.[185] The International Court of Justice (ICJ) argued that the ICTY's approach "stretches too far, almost to the breaking point, the connection which must exist between the conduct of a State's organs and its international responsibility."[186] Thus, except in perhaps cases of extensive state control over the activities of the intermediary, it is unlikely that this standard would be met in most instances of platform regulation.

It is more plausible, however, to suggest that an intermediary's activities could be considered state action when the state instructs or directs the intermediary to carry out the wrongdoing. For example, in the context of cyberattacks, Mačák argues that the test for instruction or direction might be satisfied "if a State specifically instructed an IT department within a university to carry out a Distributed Denial of Service

180. *Articles on State Responsibility*, *supra* note 167, art. 8.

181. *Report of the International Law Commission on the Work of Its Fifty-Third Session*, 56 U.N. GAOR Supp. No. 10, U.N. Doc. A/56/10 (2001), *reprinted in* [2001] 2 Y.B. INT'L L. COMM'N 1, 47, U.N. Doc. A/CN.4/SER.A/2001/Add.1 [hereinafter *Commentaries*].

182. Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. Rep. 14, ¶ 115 (June 27); *see also Commentaries*, *supra* note 181, at 47 (noting the high bar required to prove that the conduct was carried out under the direction or control of the state).

183. Hessbruegge, *supra* note 169, at 272.

184. In Tadić, the ICTY found the existence of an international conflict because the Federal Republic of Serbia exercised overall control over the activities of Serbian armed forces in Bosnia. Prosecutor v. Tadić, Case No. IT-94-1-A, Judgment, ¶¶ 131-140 (Int'l Crim. Trib. for the Former Yugoslavia, July 15, 1999). This test is less stringent, requiring only "support and co-ordination." *Id.*

185. Application of the Convention on the Prevention and Punishment of the Crime of Genocide (Bosn. & Herz. v. Serb. & Montenegro), Judgment, 2007 I.C.J. Rep. 42, ¶ 406 (Feb. 26).

186. *Id.*

(DDoS) attack against a designated target."[187] A state order directing an intermediary to remove content might similarly transform the resulting removal into an act of state triggering human rights responsibility.

In most such instances, resort to principles of state responsibility would not be necessary to ensure accountability. Although attribution is needed in the context envisioned by Mačák in order to determine whether the rules on use of force have been triggered, in cases of content removal done pursuant to the order of a state, the state is directly accountable under human rights law for the instruction or direction it issues. Government activities that involve the issuance of commands directed to an intermediary are themselves state action that must comply with human rights law, even if those activities seek to use private actors for their implementation. If a government is responsible for content removal through an injunction, law enforcement order, or even informal pressure, its actions can be evaluated directly under human rights law.[188]

This is consistent with the approach of the ICJ in *Nicaragua*. Even though the ICJ imposed a high bar for attributing state responsibility based on private action, it nonetheless easily found the state responsible for its own wrongful acts.[189] In *Nicaragua*, the United States had violated international law by providing the Contras with a manual of psychological operations that advocated violence to achieve propagandistic effects.[190] Thus, although the United States' level of control over the Contras was not enough to render it responsible for their human rights violations, the U.S. government could still be held responsible for its own acts that contributed to violations.[191]

Like direct injunctions or decrees, the actions of internet referral units must also be evaluated under human rights law. Even though these content referrals are seeking to leverage implementation by relying on private actors for actual removal, these referrals are themselves state actions that can interfere with freedom of expression. These units are state organs charged with scouring the internet to identify content the state believes is problematic and then notifying the platform that this content may violate the platform's terms of service. These referrals must comply with the requirement that limits on expression be legitimate, lawful, and

187. Mačák, *supra* note 169, at 415.

188. *See* Shaffer Van Houweling, *supra* note 5; *see also* Backpage.com, LLC v. Dart, 807 F.3d 229 (7th Cir. 2015) (law enforcement officer violated the First Amendment when he pressured credit card companies to cease serving a website based on lawful conduct).

189. Military and Paramilitary Activities in and Against Nicaragua (Nicar. v. U.S.), Judgment, 1986 I.C.J. Rep. 14, ¶ 116 (June 27).

190. *Id.* ¶ 118.

191. *Commentaries*, *supra* note 181, at 47 (even if the state does not exercise sufficient control over the intermediary such that the intermediary's wrongful acts can be attributed to the state, the state is still responsible for its own actions).

proportional. Some internet referral units already take steps to comply with this obligation by obtaining preliminary determinations of illegality before referring content to an intermediary.[192]

To some extent, it might seem that simply suggesting content for takedown would not necessarily interfere with freedom of expression.[193] As one report noted, the policy of IRUs "is not to explicitly order the removal of such content, but rather to notify the intermediary who must then independently determine whether it constitutes a breach of its Terms of Service."[194] Currently, however, these referrals are operating as much more than simply suggestions. These relationships are backed up by threats and sanctions of varying levels of formality. But even in the absence of explicit sanctions, a request from a government official in the current environment carries weight and can interfere with rights. As Brian Chang explains:

> While individual referrals may take the form of voluntary requests, IRU referrals cannot be divorced from the broader context in which they are made. The threat of excessive intermediary liability (for content that is not produced or modified by the ICT companies), the potential that their service may become blocked, and other coercive pressures mean that ICT companies have found—and will continue to find—that it makes sense commercially to err on the side of over-censorship.[195]

The result, as Daphne Keller notes, is a strong incentive to "accept all Referrals – even if that requires changing how they interpret their TOS, and taking down previously permitted expression. That means accommodating even the most aggressive Referrals from national authorities, letting their requests shape online information access throughout the EU and around the world."[196] Thus, in the current environment—particularly with the United Kingdom and the European Union poised to enact significant new regulations for internet intermediaries—these referrals should be viewed as state actions that must comply with human rights law.[197]

Some of the extra-legal partnerships that states are establishing with

192. Pielemeier & Sheehy, *supra* note 73.

193. Land, *supra* note 37, at 301.

194. *IViR Study*, *supra* note 176, at 59.

195. Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114, 122-23 (2018).

196. Keller, *supra* note 76.

197. For example, the enactment of the German NetzDG was prompted at least in part by perceptions among German officials that Facebook was not doing enough to enforce its own rules on hate speech. KAYE, *supra* note 40, at 66-68.

intermediaries also involve state action for which the state can be held directly accountable. For example, although the UK Council on Internet Safety is described as a "forum," it is in fact an initiative run by an organ of the state. It is described as "part of" three different government agencies, including the UK Home Office.[198] The Executive Board is composed of tech companies and civil society organizations but also a range of governmental entities such as GCHQ (the UK Government Communications Headquarters), the UK's intelligence and security agency, as well as representatives from law enforcement; and it is jointly chaired by three UK Ministers.[199] In the current environment, suggestions emanating from the government agencies that sit on this Council are likely to be viewed as actions needed to avoid further liability and regulation and therefore should be assessed under human rights law.

2. Intermediary Liability

The second category of government regulation of online speech consists of regimes of intermediary liability that, to varying degrees, obligate intermediaries to police and make decisions about the legality of the user speech that appears on their platforms. As described in the next Part, intermediary liability that is accompanied by appropriate safeguards to avoid overburdening speech is not clearly unlawful. But, it remains an open question as to whether content regulation done to satisfy the requirements of an intermediary liability regime must comply with the requirements of legality, legitimacy, and proportionality—i.e., whether regulation by the intermediary mandated by the state is state action and thus governed by human rights law.[200]

Like state orders, the delegation itself must of course also be judged by international human rights standards. Nonetheless, determining whether the resulting conduct of the intermediary should be evaluated under human rights law is also important. This is because the delegation alone may not provide a sufficient basis to determine its consistency with human rights law. The very nature of an intermediary liability regime requires the intermediary to decide who can speak and when. For example, a hate speech law that did not include a definition of hate speech would leave it to the intermediary to determine what speech meets that standard. Holding

198. *UK Council for Internet Safety*, https://www.gov.uk/government/organisations/uk-council-for-internet-safety.

199. Press Release, UK Dep't for Digital, Culture, Media & Sport, Board Announced for New UK Council for Internet Safety (Oct. 20, 2018), https://www.gov.uk/government/news/board-announced-for-new-uk-council-for-internet-safety.

200. Jørgensen & Pedersen, *supra* note 36, at 187 ("[A]t what stage may freedom of expression limitations be attributed to the state, when caused by 'voluntary' measures taken by intermediaries following state encouragement to do so?").

the state responsible for the delegation does not ensure accountability for how that delegation is implemented.

Under Article 5 of the International Law Commission's Articles on State Responsibility, states incur responsibility for the wrongful acts of entities, which, although not officially organs of the state, are "empowered by the law of that State to exercise elements of governmental authority."[201] Article 5 addresses entities that exercise state authority on behalf of state agencies as well as "former State corporations [that] have been privatized but retain certain public or regulatory functions."[202] What is key is that the entity "is empowered by the law of the State to exercise functions of a public character normally reserved by State organs."[203] This might include, for example, a private airline that is delegated authority over immigration issues, or a private company that is charged with running a prison and exercising the governmental power of detention.[204]

When governments delegate authority to intermediaries to regulate the speech on their platforms through systems of intermediary liability, the resulting actions by the intermediary should be considered state action under Article 5.[205] Placing legal responsibility on the intermediary to regulate the content on its platform thrusts the intermediary into the role of state censor. This delegated authority involves states regulating intermediaries not as ends, but as means—enlisting private actors to exercise regulatory discretion regarding the expression of their users.[206]

These regimes empower intermediaries to exercise authority normally reserved to the state in two primary ways. First, they are asking intermediaries to regulate the conduct of others—a kind of policing function hidden under the guise of "self-regulation." As European Digital Rights observes, states are coercing intermediaries "to police and punish their own consumers . . . under the flag of 'self-regulation' even though it is not regulation—it is policing—and it is not 'self-' because it is their consumers and not themselves that are being policed."[207] Thus, the UK White Paper's "duty of care" is not a true duty of care, but instead a duty to regulate. As Global Partners has argued, a true duty of care is a duty that exists "in relation to harm that is caused by the individual or entity's acts

201. *Articles on State Responsibility*, *supra* note 167, art. 5.

202. *Commentaries*, *supra* note 181, at 42.

203. *Id.* at 43.

204. *Id.*

205. *See* Keller, *supra* note 4, at 296 ("When OSPs remove user expression based on actual or perceived legal requirements, the harm to the user's rights can be traced to state action through laws which create OSP liability.").

206. *See generally* Kreimer, *supra* note 44, at 14 (noting "governments have sought to enlist private actors within the chain as proxy censors to control the flow of information").

207. EUROPEAN DIGITAL RIGHTS, THE SLIDE FROM "SELF-REGULATION" TO CORPORATE CENSORSHIP 1 (2011).

or omissions."[208] The duty of care proposed by the White Paper, however, is "wholly different . . . because it would make online platforms liable for the actions of third parties."[209] Policing the conduct of others in this way is a law enforcement function and constitutes an essential government activity that should not be delegated to a private entity without a framework of accountability.[210]

Second, intermediary liability regimes for harmful speech delegate essential government authority by requiring intermediaries to create law. As Bloch-Wehba notes:

> [R]ather than simply compelling intermediaries to delete specific content, governments are foisting upon platforms increasing responsibility for making legal determinations regarding speech—a task that might previously have belonged to a court, administrative agency, or other government body accountable to the public. As a result, intermediaries bear increasing duties to make important decisions regarding sensitive civil liberties issues.[211]

This kind of delegation, called co-regulation, is a "form of governance for public authorities, based on the voluntary delegation or transfer to private actors of the burden of all or part of the drafting, implementation and enforcement of norms."[212] In the context of internet regulation, governments engage in this form of governance by imposing broad, open-ended requirements on intermediaries to establish the permissible boundaries of speech. Belli and Zingales, for example, focus on open-ended injunctions directed at content removal, the right to be forgotten, and the European Code of Conduct. They note that the lack of specific guidance provided by the European Union regarding the right to be forgotten grants platforms "wide discretion in the implementation of the

---

208. GLOBAL PARTNERS DIGITAL, ONLINE HARMS WHITE PAPER CONSULTATION: GLOBAL PARTNERS DIGITAL SUBMISSION 10-11 (2019).

209. *Id.* Thus, according to Global Partners Digital: "To hold online platforms liable simply because content is generated or shared via their platforms, or users act in a way which is illegal or harmful, would run entirely contrary to existing duties of care and create an inconsistent system of liability." *Id.*

210. Luca Belli & Nicolo Zingales, *Law of the Land or Law of the Platform? Beware of the Privatisation of Regulation and Police*, *in* PLATFORM REGULATIONS: HOW PLATFORMS ARE REGULATED AND HOW THEY REGULATE US 41, 42 (Luca Belli & Nicolo Zingales eds., 2017) (arguing that these are "regulation and police functions that have traditionally been considered a matter of public law"); *see also* KAYE, *supra* note 40, at 83 (describing the proposed European Commission regulation as "asking companies to make legal decisions").

211. Bloch-Wehba, *supra* note 2, at 31-32.

212. Benoît Frydman et al., *Co-Regulation and the Rule of Law*, *in* GOVERNANCE, REGULATION AND POWERS ON THE INTERNET 133, 134 (Eric Brousseau et al. eds., 2012) (internal citation omitted).

right."[213]

In each of the examples considered by Belli and Zingales, the power granted to intermediaries "morph[s] into private lawmaking and adjudication, where platforms not only choose the means of implementation of the delegated functions, but also substantially take part in the definition and interpretation of the rights and obligations of their users."[214] Indeed, given the importance of context in assessing the lawfulness of speech,[215] it may be that most regulation of speech is necessarily an exercise in lawmaking authority. Distinguishing between extremist content and parody about extremist content, for example, requires complex and nuanced assessments about meaning and interpretation. In such contexts, applying law is equivalent to creating it.

Intermediary liability regimes also delegate the authority to create law by asking intermediaries to translate legal principles established for offline context into the online environment. It is a truism by now that the same rights that individuals enjoy offline should also be enjoyed online. Nonetheless, the way in which these rights are enforced online as opposed to offline creates substantially different impacts on speech and thus changes the extent to which a particular law may or may not be proportional to the government's objectives. Similarly, a law drafted for the offline context, if simply incorporated into content policies without additional clarification, may not provide sufficient notice to users about what content they can post. The German NetzDG law is instructive in both respects. The law itself does not create substantive categories of illegality, but requires intermediaries to enforce already established laws prohibiting, among other things, twenty-two provisions of the German Criminal Code.[216] These provisions include the crimes of "incitement to hatred" and "dissemination of depictions of violence," as well as "insult," "defamation," and "defamation of religions, religious and ideological associations in a manner that is capable of disturbing the public peace."[217]

The categories of wrongful speech that intermediaries in Germany are now expected to enforce are not only broad and ill-defined, but in practice, they also rely on considerable under-enforcement in order to be compatible with freedom of expression. To the extent that a crime such as "incitement to hatred" is enforced in the offline context, there is a significant amount of speech that inevitably escapes enforcement. An

213. Belli & Zingales, *supra* note 210, at 51.

214. *Id.* at 47.

215. Richard Ashby Wilson & Molly K. Land, *Hate Speech on Social Media: Towards a Context-Specific Content Moderation Policy*, 52 CONN. L. REV. 1, 30-40 (2020).

216. Heidi Tworek & Paddy Leerssen, *An Analysis of Germany's NetzDG Law* 2 (Transatlantic Working Grp., 2019).

217. *Id.*

individual who incites hatred in a comment directed to a group of friends, or even one who shouts an invective at someone in public, will likely simply be ignored by the state—and, indeed, that is important, since perfect enforcement of such a provision would cut deeply into freedom of expression. Given the difficulty of perfect enforcement, state sanction ends up being applied to those instances in which the speech is particularly dangerous in light of the number of people it reaches, the context in which it occurs, or the influence of the speaker.[218]

There is considerable speech that escapes enforcement in the online context as well. Indeed, this has been a driving force behind increased state pressure on intermediaries. However, the *distribution* of non-enforcement online is fundamentally different than the distribution of non-enforcement in the offline context. In the offline context, non-enforcement of speech-related rules generally accounts for a range of factors that make speech more or less dangerous. This is done through the inclusion of standards that require consideration of context,[219] but also through discretionary determinations by law enforcement and prosecutors about which wrongful acts to pursue. In the online context, however, the distribution of non-enforcement looks very different. Because of the challenges of enforcement at scale, enforcement in the online context has tended to focus on the content of the speech, rather than the (often far more important) non-content factors that make speech more or less risky. The result is that enforcement of speech online is simultaneously overbroad and underinclusive when compared to its offline equivalent.

Lacking direction from states on how to translate offline rules into online spaces, companies have created what David Kaye has called "platform law."[220] Platform law includes not just contractual provisions embodied in a platform's terms of service but also community standards, content moderation practices and decisions, and internal guidance provided to employees.[221] This substantive law is coupled with extensive procedural law about how content is flagged and reviewed as well as technical law that specifies the way in which content can be posted, viewed, shared, and moderated.[222] Indeed, the very creation of such robust legal systems is convincing evidence of the governmental nature of the

218. *See generally* RICHARD ASHBY WILSON, INCITEMENT ON TRIAL: PROSECUTING INTERNATIONAL SPEECH CRIMES (2017).

219. *See, e.g.*, Brandenburg v. Ohio, 395 U.S. 444 (1965).

220. Kaye, *supra* note 36, ¶ 1.

221. Molly K. Land, *The Problem of Platform Law: Pluralistic Legal Ordering and Social Media*, *in* OXFORD HANDBOOK ON GLOBAL LEGAL PLURALISM 958, 964-65 (Paul Berman ed., 2020); *see also* Kaye, *supra* note 36, ¶ 63 (calling on companies to make their decisions about content available to the public as a kind of caselaw that can help communicate to users what kind of speech is not allowed on their platforms).

222. Land, *supra* note 221, at 964-65.

authority that these companies have been delegated.

Unlike speech rules in the offline context, platform law ends up focusing not on the impact of speech, but on its content. This is understandable, to some extent. Platforms are struggling with how to operationalize fact-specific and discretionary rules on a global scale as applied to vast amounts of content. The result, unfortunately, is a "one-size-fits-all" platform law that—at least in terms of how it is drafted—treats as equivalents a racial slur used in a casual conversation online and a government-backed genocidal campaign of online denigration against a vulnerable ethnic minority.[223] The result is that platforms end up policing some speech far more heavily than is needed, and letting slide other speech that—while perhaps less suspect on its face—has potentially far greater consequences. Further, as automation advances, enforcement online will become even less leaky, sweeping up many instances of speech that would have gone unnoticed by the state in the offline context. These questions about the distribution of enforcement substantially alter the proportionality of speech restrictions and are questions that invokes the authority of the state. Thus, delegation of censorship authority to platforms not only requires the platform to make determinations about what the law is, but also about how it should be enforced.

Of course, this does not mean that imposing liability on corporate actors necessarily transforms their activities into state action. Clearly, a state can and should when appropriate impose liability on private actors, including gatekeepers.[224] The secondary liability of a bookstore for defamation also requires the store to make legal determinations. But secondary liability for defamation is distinct from intermediary liability because the latter requires the intermediary to go beyond regulating its own behavior (i.e., whether to sell a particular book) to creating a regulatory regime for its users that determines who may speak and when. Further, the impact on expressive rights of the largest platforms is much more significant than any individual bookstore. Finally, liability for defamation involves qualitatively different considerations when applied online and at scale, considerations that themselves require the exercise of lawmaking authority to translate offline concepts into the online context. When the delegated responsibility includes regulatory and lawmaking functions such as these, the resulting activity is state action which must comply with human rights standards.

3. Extra-Legal Influence

Techniques of extra-legal influence can, depending on the

223. Land & Hamilton, *supra* note 55, at 144.
224. *See* Kraakman, *supra* note 147, at 53-54.

circumstances, be judged with reference to one or both of the categories discussed above. To the extent that the extra-legal influence is pressure from a government official that results in removal of content, the action of the state agent can be evaluated under human rights law. Further, where systems of voluntary self-regulation are actually delegated authority in disguise, it could constitute state action under Article 5. Although heralded as "voluntary self-regulation," for example, the EU Code of Conduct was not really voluntary. EU regulators made clear that the only alternative to such a code was state regulation.[225] As evidence of its coercive power, the Code was even negotiated together with governments, thus embodying an agreement—albeit non-binding—about the practices of the platforms. Finally, the Code was also highly specific about the kind of content that must be removed and the time frame for removal,[226] reinforcing the conclusion that it represented a disguised governmental command. Thus, while informal pressure would not ordinarily rise to the level of delegation, coupling this pressure with threats of sanction, particularly in an environment in which those threats are highly credible, means the resulting regulation by the intermediary should comply with human rights standards.

IV. A HUMAN RIGHTS NON-DELEGATION DOCTRINE

This Part considers whether "collateral"[227] or "delegated"[228] censorship is permissible under international human rights law. It argues that such censorship—which today is poised to become the leading form of regulatory control over online content—is unlawful under international human rights law unless accompanied by meaningful safeguards to ensure accountability.[229] Although governments have a positive duty to create a regulatory environment in which all users' rights can be respected, online as well as offline,[230] including the obligation to protect individuals from the very real harms of online speech,[231] states cannot fulfill this obligation by

225. Keller notes, for example, that the agreement "was widely perceived as a compromise to stave off legislation." Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money* 8 (Hoover Working Grp. on Nat'l Sec., Tech., and Law, Aegis Series Paper No. 1807, 2018), https://lawfareblog.com/internet-platforms-observations-speech-danger-and-money.

226. Evelyn Aswad, *The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban*, 1 COLUM. HUM. RTS. L. REV. ONLINE 1, 3-4 (2016).

227. Balkin, *supra* note 3, at 2298; Christina Mulligan, *Technological Intermediaries and Freedom of the Press*, 66 SMU L. REV. 157, 164-65 (2013).

228. DENARDIS, *supra* note 6, at 158; Bambauer, *supra* note 8, at 879.

229. Keller asks a similar question under U.S. law, concluding—based on earlier cases addressing the liability of bookstores—that intermediary liability may be inconsistent with the First Amendment due to the impact of collateral censorship. *See* Keller, *supra* note 225, at 16-20.

230. AGUSTINA DEL CAMPO, CONTENT MODERATION AND PRIVATE CENSORSHIP: STANDARDS DRAWN FROM THE JURISPRUDENCE OF THE INTER-AMERICAN HUMAN RIGHTS SYSTEM 5-6 (2017).

231. La Rue, *supra* note 163, ¶¶ 24-33.

simply shifting it to internet intermediaries, nor can they use intermediaries to pursue state aims under the guise of private action.[232]

A. Delegated Censorship Under Human Rights Law

Delegated censorship unaccompanied by safeguards is unlawful under international human rights law because it is inevitably overbroad. Under international human rights law, limits on speech are permissible, but any such limits must be "provided by law and [be] . . . necessary: (a) For respect of the rights or reputations of others; (b) For the protection of national security or of public order (ordre public), or of public health or morals."[233] Thus, any limitations on speech "must be provided by law, which is clear and accessible to everyone (principles of predictability and transparency)," "must pursue one of the purposes set out in Article 19, paragraph 3, of the Covenant . . . (principle of legitimacy)," and "must be proven necessary and the least restrictive means required to achieve the purported aim (principles of necessity and proportionality)."[234]

Without limits and safeguards, censorship by platforms will be overbroad and thus not proportional to the state's objectives. First, privatized censorship will be disproportionate because of the lack of alignment between the incentives of the platform and the speaker.[235] Internet companies do have an interest in protecting freedom of expression on their platforms as a general matter, since the content that users generate is an essential component of the product that they provide. But, while platforms may have an interest in protecting freedom of expression generally, they do not necessarily have a strong interest in protecting any particular individual instance of speech. Thus, as Meléndez-Juarbe explains:

> To the extent that intermediaries and users have divergent interests, a given intermediary will not necessarily take into account the value that the regulated activity has to the user. Instead, the intermediary will rationally behave so as to maximize its welfare, seeking to minimize its expected liability cost. In the end the problem is that, when balancing private liability costs

232. MANFRED NOWAK, HUMAN RIGHTS OR GLOBAL CAPITALISM: THE LIMITS OF PRIVATIZATION 132 (2017) ("But that the state cannot avoid its responsibility by delegating its functions to non-state actors is fairly uncontested under both international and domestic law in most states.").

233. International Covenant on Civil and Political Rights art. 19(3), Mar. 23, 1976, 999 U.N.T.S. 171.

234. La Rue, *supra* note 163, ¶ 24.

235. Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 NOTRE DAME L. REV. 293 (2013); *see also* KAYE, *supra* note 40, at 70; Kreimer, *supra* note 44, at 28-29.

> against the private benefits of an intermediary's economic activity, in its self-interested conception of what constitutes wellbeing, the intermediary will not consider user's free speech interests (or the social interest in that the user engages in activity presumably connected to freedom of expression).[236]

The costs of taking the time to ensure they are protecting freedom of expression are very high for intermediaries, given the volume of speech they must manage and the short time frame in which to assess it.[237] Further, because intermediaries lack the incentive to investigate claims, they are also vulnerable to third parties who might seek to game liability regimes to achieve other purposes.[238]

The cost of liability is also increasing significantly. The NetzDG imposes fines of up to fifty million Euro per violation for failure to remove manifestly unlawful content within 24 hours.[239] An intermediary's failure to comply with the proposed new law in the United Kingdom could result not only in fines but also individual criminal and civil liability for the company's senior management.[240] Faced with the high cost of liability and lacking investment of their own in the expression at issue,[241] intermediaries are likely to suppress more speech than they would if it were their own.[242] Those whose speech is incorrectly removed can in some instances challenge the removal, but that requires that they know their speech has been removed and that they have the motivation and resources needed to challenge the takedown.[243]

Empirical evidence backs up these concerns about the suppression of speech by intermediaries charged with policing their users. Urban and Quilter's 2005 study of copyright takedown notices under Section 512 of the Digital Millennium Copyright Act found that the process was commonly used for purposes other than copyright protection, including "to create leverage in a competitive marketplace, to protect rights not given by copyright (or perhaps any other law), and to stifle criticism,

---

236. Meléndez-Juarbe, *supra* note 79, at 111.

237. Wu, *supra* note 235, at 307. They are also less likely than speakers to assume the lawfulness of their speech. *Id.* (noting that speakers suffer from confirmation bias).

238. Kreimer, *supra* note 44, at 32; *see also* Ciolli, *supra* note 11, at 184-85 (discussing the chilling effect of frivolous lawsuits against immunized intermediaries).

239. *Germany: Flawed Social Media Law—NetzDG Is Wrong Response to Online Abuse*, HUM. RTS. WATCH (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

240. ONLINE HARMS WHITE PAPER, *supra* note 48, at 7; *see also id.* at 60.

241. Wu discusses a range of benefits that speakers might enjoy from their speech that are not also enjoyed by intermediaries, including monetary or reputational benefits, revenge, social obligation, and self-expression. Wu, *supra* note 235, at 304-06.

242. Kreimer, *supra* note 44, at 27; Balkin, *supra* note 3, at 2309.

243. Balkin, *supra* note 3, at 2314.

commentary and fair use."[244] In particular, they found that "a large number of notices present[ed] serious substantive questions about the underlying claim" and that takedown occurred nonetheless "in numerous questionable situations."[245] Bar-Ziv and Elkin-Koren's study of removal requests sent to Google regarding content on Israeli websites that allegedly infringed copyright revealed that sixty-six percent of the requests had "little to do with copyright."[246] Townend's interviews with community news bloggers and journalists in England and Wales revealed a spectrum of responses to threats of libel and piracy; while some were unaffected by the threat of such risk, others revealed "excessive self-censorship."[247]

Second, censorship by intermediaries is also not transparent or predictable. Limits on speech must comply with the principle of legality, which requires regulators to convey information about what conduct is permitted and what is not. Regulation by intermediaries, however, is less visible, since decisions about content are made by private actors behind closed doors. Companies may be subject to competitive or legal constraints on what they can share, and their decisions are often implemented through technical means that may not be transparent.[248] Few companies provide information about whether and if so to what extent they engage in filtering or other content restriction at the request of governments.[249] (And governments are also not transparent about the nature and scope of the demands they put on companies.[250]) Individuals

244. Jennifer M. Urban & Laura Quilter, *Efficient Process or Chilling Effects—Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 SANTA CLARA HIGH TECH. L.J. 621, 687 (2005). Research published by Urban, Karaganis, and Schofield confirmed that this trend continues. They note that 31% of the takedown requests that they analyzed were found to be "potentially problematic," including over 4% that were "fundamentally flawed because they targeted content that did not clearly match the identified infringed work." JENNIFER M. URBAN ET AL., NOTICE AND TAKEDOWN IN EVERYDAY PRACTICE 11 (2017); *see also* RISHABH DARA, INTERMEDIARY LIABILITY IN INDIA: CHILLING EFFECTS ON FREE EXPRESSION ON THE INTERNET 2 (2011) (finding that "[o]f the 7 intermediaries to which takedown notices were sent, 6 intermediaries over-complied with the notices, despite the apparent flaws in them"); BRENNAN CTR. FOR JUSTICE AT NYU SCH. OF LAW, WILL FAIR USE SURVIVE? FREE EXPRESSION IN THE AGE OF COPYRIGHT CONTROL: A PUBLIC POLICY REPORT ii (2005) (finding, based on an "analysis of 320 cease and desist and take-down letters from the Chilling Effects Web site" that "more than 20% either stated weak copyright or trademark claims, or involved speech with a strong or at least reasonable free expression or fair use defense" and that "[a]nother 27% attacked material with possible free expression or fair use defenses").

245. Urban & Quilter, *supra* note 244, at 681.

246. Sharon Bar-Ziv & Niva Elkin-Koren, *Behind the Scenes of Online Copyright Enforcement: Empirical Evidence on Notice & Takedown*, 50 CONN. L. REV. 1, 5 (2017).

247. Judith Townend, *Online Chilling Effects in England and Wales*, 3 INTERNET POL'Y REV. 1, 2 (2014).

248. Balkin, *supra* note 3, at 2297-98; Kreimer, *supra* note 44, at 28.

249. *Corporate Accountability Index*, RANKING DIGITAL RTS., https://rankingdigitalrights.org/index2019/indicators/f6/.

250. KAYE, *supra* note 40, at 123-24.

whose content has been removed have few if any opportunities to challenge that removal, even if they have the resources to do so.[251]

Third, broad delegations of censorship authority also involve transfer of an inherent governmental function—lawmaking authority—that is inconsistent with human rights law. Although privatization is compatible with human rights law and can be an effective and efficient way of realizing important governmental objectives, at least some kinds of core governmental functions "cannot be outsourced to the corporate sector without violating human rights."[252] As Nowak explains, while privatization might be permissible even when it raises human rights concerns, there are other kinds of delegation, "the implementation of which requires states to take specific measures which are generally considered inherent government functions."[253] Nowak identifies a middle category of privatization that involves the private sector "in areas that have traditionally been considered inherent public functions and which have a direct impact on the enjoyment of certain human rights."[254] In this category, Nowak distinguishes between "mere support functions and services directly related to the enjoyment of human rights."[255] Core functions, he argues, cannot be delegated. As an example, he argues that internal and external security functions are "inherent governmental functions that may not be privatized or outsourced" without violating the individual right to personal security.[256]

Limiting speech to protect the rights of others and to achieve important public policy goals is an inherent governmental function. Requiring a private actor to determine, especially without guidance from the state, whether expressive content online is lawful satire or impermissible incitement, is equivalent to the inherent functions governments exercise in ensuring the right to vote or the right to a fair trial—two examples that Nowak notes may not be privatized.[257] As Karanicolas argues:

> Regulating speech is among the most important, and most delicate, tasks that a government may undertake. It requires a careful balancing between removing harmful content while providing space for controversial and challenging ideas to spread,

251. *See* Kreimer, *supra* note 44, at 27, 31.
252. NOWAK, *supra* note 232, at 165.
253. *Id.*
254. *Id.* at 166.
255. *Id.* at 167.
256. *Id.* at 164.
257. *Id.* at 165-66.

> and between deterring dangerous speech while minimizing a broader chilling effect that can impact legitimate areas of debate.[258]

In light of the risks associated with delegation of such a core governmental function, the Human Rights Committee, the UN body responsible for receiving state reports regarding compliance with the ICCPR, has emphasized: "A law may not confer unfettered discretion for the restriction of freedom of expression on those charged with its execution."[259] Respecting freedom of expression requires states to establish parameters for delegation that comport with the principles of legality, legitimacy, and proportionality.

Given the range of concerns associated with delegated censorship authority, United Nations and regional human rights experts have consistently expressed concern about intermediary liability, including both the current and former UN Special Rapporteurs on the Promotion and Protection of Freedom of Opinion and Expression.[260] In his 2011 report, Frank La Rue, the former UN Special Rapporteur, went so far as to say that "censorship measures should never be delegated to a private entity, and . . . no one should be held liable for content on the Internet of which they are not the author."[261] Imposing liability on intermediaries for content created and shared by others, he explained, "severely undermines the enjoyment of the right to freedom of opinion and expression, because it leads to self-protective and over-broad private censorship, often without transparency and the due process of the law."[262] La Rue reiterated this position in his 2012 report, arguing that "[s]tates should request the removal of content only through a court order and intermediaries should never be held liable for content of which they are not the authors."[263]

A 2011 Joint Declaration authored by La Rue together with regional human rights experts from the Organization for Security and Cooperation in Europe, the Organization of American States, and the African Commission on Human and Peoples' Rights, also argued that intermediary liability is inherently inconsistent with human rights law:

> No one who simply provides technical Internet services such as providing access, or searching for, or transmission or caching of

258. Michael Karanicolas, *Privatizing Censorship*, BALKANIZATION (June 12, 2019), https://balkin.blogspot.com/2019/06/privatizing-censorship.html.

259. Human Rights Comm., General Comment No. 34: Article 19: Freedom of Opinion and Expression, ¶ 25, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011).

260. Kaye, *supra* note 36, ¶¶ 40-41; La Rue, *supra* note 163, ¶¶ 38-48.

261. La Rue, *supra* note 163, ¶ 43.

262. *Id.* ¶ 40.

263. Frank La Rue, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 87, U.N. Doc. A/67/357 (Sept. 7, 2012).

> information, should be liable for content generated by others, which is disseminated using those services, as long as they do not specifically intervene in that content or refuse to obey a court order to remove that content, where they have the capacity to do so ('mere conduit principle').[264]

Thus, the 2011 Joint Declaration argues that "[c]ontent filtering systems which are imposed by a government or commercial service provider and which are not end-user controlled are a form of prior censorship and are not justifiable as a restriction on freedom of expression."[265] The Joint Declaration recommends that "[a]t a minimum, intermediaries should not be required to monitor user-generated content and should not be subject to extrajudicial content takedown rules which fail to provide sufficient protection for freedom of expression (which is the case with many of the 'notice and takedown' rules currently applied)."[266]

Other human rights authorities have also condemned intermediary liability. The Office of the Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights has argued that strict liability and notice and takedown are both inconsistent with the American Convention on Human Rights.[267] And in 2011, the Human Rights Committee noted that states must ensure that any restrictions placed on intermediaries comply with the limitations that Article 19 of the ICCPR puts on any restrictions on a right—namely, that the limitation be provided by law, imposed for a legitimate purpose, and conform to the tests of necessity and proportionality.[268]

The European Court of Human Rights, however, has been somewhat more ambiguous in its response to intermediary liability, holding in one case that a government may impose penalties on a media site for failing to proactively police the content on its webpage. In *Delfi v. Estonia*, the Grand Chamber of the European Court of Human Rights (ECtHR) addressed a petition brought by Delfi, an internet news portal, against the government of Estonia alleging that Estonia's decision to hold it responsible for defamatory content posted by readers in the comments section under an article on its website violated Delfi's rights to freedom of expression under Article 10 of the European Convention on Human Rights (ECHR).[269]

---

264. Org. for Sec. & Cooperation in Eur. [OCSE], *Joint Declaration on Freedom of Expression and the Internet*, ¶ 2(a) (2011) [hereinafter *Joint Declaration*].

265. *Id.* ¶ 3(b).

266. *Id.* ¶ 2(b).

267. CATALINA BOTERO MARINO, INTER-AMERICAN COMM'N ON HUM. RIGHTS, FREEDOM OF EXPRESSION AND THE INTERNET ¶¶ 98, 105 (2013).

268. Human Rights Comm., *supra* note 259, ¶¶ 22, 43.

269. Delfi AS v. Estonia, 2015-II Eur. Ct. H.R. 319, ¶ 16.

Delfi had published a news article about the destruction of ice roads between the Estonian mainland and certain islands. In the comments section, users had posted comments that included defamation of and personal threats directed to the ferry owner allegedly responsible.[270] Although Delfi removed the comments quickly after being notified of their presence, Estonia still fined the news portal for the time the posts were on the site prior to notice.[271] The ECtHR found that the imposition of this fine was not inconsistent with Delfi's rights to freedom of expression under the ECHR. Given the risks that harmful speech can pose for individual rights, the ECtHR found that member states could appropriately "impose liability on Internet news portals, without contravening Article 10 of the Convention, if they fail to take measures to remove clearly unlawful comments without delay, even without notice from the alleged victim or from third parties."[272]

A subsequent case by the Fourth Section of the ECtHR, however, distinguished and appeared to limit the holding in *Delfi* to impose an affirmative monitoring duty only for particularly noxious content. In *Case of Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v. Hungary*, the Fourth Section considered whether Hungary could impose penalties on a website for not proactively removing defamatory comments on their sites regarding the alleged unethical conduct of two real estate developers.[273] Using factors identified in *Delfi*, the Fourth Section distinguished the case before it based on the relative harmlessness of the comments, noting that the comments at issue, "[a]lthough offensive and vulgar . . . , did not constitute clearly unlawful speech; and they certainly did not amount to hate speech or incitement to violence."[274] Thus, the court narrowed the holding of *Delfi* to allow the use of notice-and-take-down "as an appropriate tool for balancing the rights and interests of all those involved," and to require proactive monitoring only when the content "take[s] the form of hate speech and direct threats to the physical integrity of individuals."[275]

B. Elements of Responsible Governance

Given the quantity of speech involved, as well as the difficulty of holding individual users accountable for the harms of their speech at scale

270. *Id.* ¶¶ 16-17.
271. *Id.* ¶¶ 27-31.
272. *Id.* ¶ 159.
273. Magyar Tartalomszolgáltatók Egyesülete v. Hungary, Eur. Ct. H.R., ¶¶ 11-14 (Feb. 2, 2016), http://hudoc.echr.coe.int/eng?i=001-160314 (2016).
274. *Id.* ¶ 64.
275. *Id.* ¶ 91.

and across borders, state regulatory efforts will inevitably involve intermediaries.[276] And, relying on private actors to provide services and carry out governmental objectives is in general not inconsistent with human rights law. Nonetheless, states must ensure that privatization does not undermine the guarantees of international human rights.[277] Safeguards must be particularly rigorous when the state has outsourced the exercise of essential governmental authority. As the Inter-American Court of Human Rights explained in *Ximenes-Lopes*:

> Though the States may delegate the rendering of such [public] services, through the so-called outsourcing, they continue being responsible for providing such public services and for protecting the public interest concerned. Delegating the performance of such services to private institutions requires as an essential element the responsibility of the States to supervise their performance in order to guarantee the effective protection of the human rights of the individuals under the jurisdiction thereof and the rendering of such services to the population on the basis of non-discrimination and as effectively as possible.[278]

Thus, to the extent that states involve intermediaries in their efforts to protect individuals from the harms of online speech, they must ensure that they minimize the negative impacts this might have on freedom of expression.[279]

This sub-Part outlines a framework for state approaches to intermediary liability that comply with the state's obligation to minimize negative impacts on freedom of expression and to ensure that any delegation is not overbroad and thus impermissible under human rights law. First, it endorses proposals that would impose liability only for some kinds of particularly harmful content but which otherwise make an intermediary liable only for its own culpable conduct. Second, it argues that any delegation must be clearly defined and provide guidance to the intermediary about how to apply the law. This guidance must also take into account the differential impact on speech that may result when laws designed for offline application are translated into the online context.

---

276. Evelyn Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation* 1 (Hoover Working Grp. on Nat'l Sec., Tech., and Law, Aegis Series Paper No. 1903, 2019) ("As the actors closest to the front line, platforms will always need to play a significant role in drawing lines for online speech, given the high-volume, fast-moving, and context-dependent nature of the decisions involved.").

277. NOWAK, *supra* note 232, at 54.

278. Ximenes-Lopes v. Brazil, Merits, Reparations, and Costs, Judgment, Inter-Am. Ct. H.R. (ser. C) No. 149, ¶ 96 (July 4, 2006).

279. Arun, *supra* note 94, at 86-87.

Third, it argues in favor of targeted accountability mechanisms and moderation by design.[280]

1. Differentiated Liability

Many of the proposals for reforming intermediary liability suggest differentiated liability based on either the nature/activities of the intermediary, or on the specific kind of content at issue. Early proposals, for example, suggested that intermediaries acting as "mere conduits" for speech should be immune from liability for that speech. Others have recommended liability based on the platform's culpability, maintaining that those who benefit from the harmful speech (or who perhaps even solicit and encourage it) should be held responsible for that speech. Content-based proposals, in contrast, advocate distinctions based on different types of speech, imposing liability for some kinds of speech but immunizing other kinds where the costs to expression would be high.

Historically, much of the critique of intermediary liability focused on the inappropriateness of placing liability on intermediaries acting as "mere conduits" for the speech of others. The 2011 Joint Declaration, for example, emphasizes that liability for "content generated by others" may not be imposed on any service that "simply provides technical Internet services," as long as the service does not "specifically intervene in that content or refuse to obey a court order to remove that content."[281] Former UN Special Rapporteur La Rue stated that intermediaries should not be liable for content "of which they are not the author."[282] The Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights also argued that holding intermediaries strictly liable for the content of others is incompatible with the guarantees of free expression.[283]

Civil society critique, as well, has emphasized the harms of liability imposed on intermediaries for the content of others. Based on the principles articulated by La Rue among others, in 2015 a coalition of civil society organizations launched the Manila Principles, a collection of guidelines based on human rights law that delineate when states may

280. This Article addresses only the question of how states should regulate intermediaries with respect to the harms associated with speech on their platforms. It does not address the broader regulatory questions of how to regulate the range of harms that might be associated with the dominance of intermediaries in the market. *See, e.g.*, K. Sabeel Rahman, *Curbing the New Corporate Power*, BOS. REV. (May 4, 2015), http://bostonreview.net/forum/k-sabeel-rahman-curbing-new-corporate-power; *see also* Lina M. Khan, Note, *Amazon's Antitrust Paradox*, 120 YALE L.J. 564 (2017); Dina Srinivasan, *The Antitrust Case Against Facebook: A Monopolist's Journey Towards Pervasive Surveillance in Spite of Consumers' Preference for Privacy*, 17 BERK. BUS. L.J. 39 (2019).

281. *Joint Declaration*, *supra* note 264, ¶ 2(a).

282. La Rue, *supra* note 163, ¶ 43.

283. BOTERO MARINO, *supra* note 267, ¶¶ 93-103.

impose liability on intermediaries.[284] Concerned about the collateral consequences of liability and in particular the burden such liability may place on speech, the Manila Principles emphasize that liability should not be imposed if the intermediary has not been involved in modifying the content.[285]

Although the protection of "mere conduit" intermediaries continues to make theoretical sense, it no longer works today as a practical matter. As Gillespie notes:

> It is not just that all platforms moderate, nor that they have to moderate, nor that they tend to disavow it while doing so. It is that moderation, far from being occasional or ancillary, is in fact an essential, constant, and definitional part of what platforms do. Moderation is the essence of platforms. It is the commodity they offer. It is their central value proposition.[286]

Thus, very few intermediaries—if any—function as "mere" conduits. Platforms sort, organize, moderate, and curate their users' content in ways designed to achieve the companies' economic goals.[287] They use proprietary formulas to decide what search results to display, what content to show, and what related content to offer, even in some instances triggering particular content in order to counter harmful messages to which the user was exposed, or experimenting with user emotions.[288] Even traditional "conduit" internet service providers like telecommunications companies can now monitor and control content due to advances in filtering technology as well as the development of techniques such as deep packet inspection.[289]

Other proposals have emphasized the culpable conduct of the intermediary itself and the extent to which the platform contributes to the harm of the speech it hosts. Danielle Citron, for example, argues in favor

---

284. *Manila Principles on Intermediary Liability: Best Practices Guidelines for Limiting Intermediary Liability for Content to Promote Freedom of Expression and Innovation*, ELEC. FRONTIER FOUND. (Mar. 23, 2015), https://www.manilaprinciples.org/[hereinafter *Manila Principles*].

285. *Id.* (Principle I(b)).

286. Tarleton Gillespie, *Platforms Are Not Intermediaries*, 2 GEO. L. TECH. REV. 198, 201 (2018).

287. Land, *supra* note 37, at 289-91.

288. *See, e.g.*, Olivia Solon, *Facebook Policy Chief: Social Media Must Step Up Fight Against Extremism*, GUARDIAN (Mar. 12, 2017), https://www.theguardian.com/culture/2017/mar/12/facebook-policy-chief-social-media-must-step-up-fight-against-extremism; Jacob Brogan, *YouTube Starts Redirecting People Who Search for Certain Keywords to Anti-Terrorism Videos*, SLATE (July 21, 2017), http://www.slate.com/blogs/future_tense/2017/07/21/youtube_redirects_those_who_search_for_terrorist_keywords_to_anti_terrorist.html; Vindu Goel, *Facebook Tinkers with Users' Emotions in News Feed Experiment, Stirring Outcry*, N.Y. TIMES (June 29, 2014), https://www.nytimes.com/2014/06/30/technology/facebook-tinkers-with-users-emotions-in-news-feed-experiment-stirring-outcry.html?_r=0.

289. Bridy, *Graduated Response*, *supra* note 9, at 123-24.

of a standard that takes into account the intermediary's business model and whether the intermediary has encouraged or profited from harmful speech. While some sites are careful to monitor for abuse, others have made a business model out of hosting and promoting such speech and often make money from extorting victims before removing harmful content.[290] Citron argues that Section 230 should be altered to make intermediaries responsible for harmful content such as nonconsensual pornography that they instigate and from which they profit.[291] Citron's approach is similar to Felix Wu's activity-based proposal, which suggests that we look at the social context to determine who is actually acting as a conduit versus a speaker.[292] In some instances, for example, an intermediary may repost the content of others but do so in a way that reflects a desire to communicate that message. In those instances, the intermediary may appropriately be held responsible as a speaker.[293]

The European Court of Human Right's decision in *Delfi* also suggests a model in which intermediaries that benefit from harmful speech are held to a higher standard. The Estonian Supreme Court emphasized, for example, that Delfi actively invited comments and benefited financially from the comments because they drove visitors to the site and thus boosted advertising.[294] The Court's decision in *Delfi* explained that liability in this instance was reasonable because of the economic benefit to the portal as well as the foreseeability of harm, the control exercised by the portal over the content, and the size, resources, and professional nature of the news portal.[295] The Fourth Section in *MTE v. Hungary* also distinguished *Delfi* on this basis, noting that one of the applicants in *MTE* was a "non-profit regulatory association of Internet service providers" that did not have economic interests in the speech at issue.[296]

Other proposals have suggested distinguishing between intermediaries based on the types of content they host or regulate. An early proposal by the Council of Europe, for example, recommended treating platforms for news differently from those directed to platforms for political debate and

290. CITRON, *supra* note 32, at 168, 173-76.

291. *Id.* at 177-81; *see also* SHIELDING THE MESSENGERS, *supra* note 14, at 11 (arguing that although the safe harbor of Section 230 plays an important role in promoting free speech and innovation, "[s]afe harbors need not . . . be readily available to clear 'bad actors' that are actively and knowingly aiding or conspiring in unlawful activity").

292. Wu, *supra* note 235, at 333-34.

293. *Id.* at 335 ("What we want to know is not whether these are someone else's facts, but whether this is someone else's message. When an entity is conveying someone else's message, that is when concerns over collateral censorship arise, and when immunity is consequently appropriate.").

294. Delfi AS v. Estonia, 2015-II Eur. Ct. H.R. 319, ¶ 31.

295. *Id.* ¶¶ 115, 117, 144-45, 153.

296. Magyar Tartalomszolgáltatók Egyesülete v. Hungary, Eur. Ct. H.R., ¶ 64 (Feb. 2, 2016), http://hudoc.echr.coe.int/eng?i=001-160314 (2016).

entertainment.[297] Of course, distinctions based on a broad "category" of activity are impracticable given how many of these functions large platforms serve simultaneously, but more tailored approaches reflect a similar sense that the content that is being regulated matters a great deal. A proposal by Meléndez-Juarbe, for example, recommends rules that "vary according to the kind of communicative act being targeted."[298] He argues:

> Thus, if we care little about the risks involved in an intermediary's efforts to prevent the circulation of child pornography, maybe we could tolerate a kind of intermediary regulatory regime that imposes stronger policing duties than the ones we would tolerate for the copyright or privacy contexts. Overenforcement in these latter two contexts presents risks of political, artistic or cultural censorship that might not be present in the former context (or at least are present to a lesser degree).[299]

Greater precautions are needed where the risks to freedom of expression are highest. For content that has low expressive value and high risk of harm, such as child pornography, we might easily tolerate greater intermediary responsibility.[300]

I propose a balanced intermediary liability regime that combines strict liability for a limited category of content that is associated with clear harm, with a presumption of immunity that can be rebutted by a showing of a lack of due diligence. First, as argued by Meléndez-Juarbe, content associated with high offline harm and low risk to freedom of expression—such as child pornography—would justify the imposition of strict liability.[301] Of course, there may still be political, artistic, or cultural censorship of content under the guise of enforcing laws against child

---

297. *Recommendation CM/Rec(2011)7 of the Committee of Ministers to Member States on a New Notion of Media*, COUNCIL EUR. 3 (Sept. 21, 2011), https://www.osce.org/odihr/101403?download=true ("For example, policy responses for media focussing on news services may differ from those offering a platform for political debate or entertainment, in turn different from the mere association of revenue-generating activities to the dissemination of content through means of mass communication.").

298. Meléndez-Juarbe, *supra* note 79, at 112.

299. *Id.* at 115.

300. Michal Lavi's proposal also reflects an effort to distinguish between intermediaries based on impact. At least with respect to speech torts, he proposes distinguishing between intermediaries based on the strength of the social ties on that platform, arguing that this determines the amount of damage associated with the tort. Michal Lavi, *Content Providers' Secondary Liability: A Social Network Perspective*, 26 FORDHAM INTELL. PROP. MEDIA & ENT. L.J. 855 (2016).

301. Meléndez-Juarbe, *supra* note 79, at 115; *see also* Keller *supra* note 225, at 19 (noting that "the state's interest in passing a law—and its tolerance for collateral damage to speech—may vary depending on the threats the law averts" and that the "kind of content at issue will also affect platforms' likely error rate and the value of procedural protections or other statutory 'tailoring' to reduce such errors").

pornography; indeed, the controversy over Facebook's removal of the iconic photograph from the Vietnam war of a naked child fleeing a napalm attack is one such example. Nonetheless, the risk of error in identifying child pornography is lower because hashing of images allows more accurate monitoring.[302]

Differentiated liability of this type has been adopted by the Argentinian Federal Supreme Court of Justice in *Rodriguez v. Google*. Columbia's Global Freedom of Expression Database summarizes the decision as follows: "According to Court, search engine companies are strictly liable for providing access to materials that clearly pose danger or harm to the public, such as child pornography or contents that facilitate or incite crimes. In comparison, they are negligently liable for contents that adversely affect one's reputation or right to privacy."[303] The Court required notice for the latter type of harmful speech, and overturned the part of the lower court's decision that imposed a monitoring obligation on the defendant intermediaries.

Although states and companies are increasingly trying to treat extremist content like child pornography,[304] there are compelling reasons to limit strict liability with a monitoring obligation only to the latter. The impact of overbroad removal of extremist content has the potential to cut more deeply into freedom of expression, and can also undermine accountability and law enforcement efforts.[305] For example, vigorous enforcement against extremist content has had negative impacts on human rights defenders who have used platforms to document and share information about human rights abuses in Syria.[306] The definition of "extremist" is also extremely broad and its application poses considerable risks of overbroad application.

For content other than child pornography, intermediaries should enjoy a general presumption of immunity. This immunity, however, can be rebutted if the intermediary's own conduct contributes to the harm of the

302. Although the costs to freedom of expression may be lower for some categories of content, there will still be privacy costs, since monitoring may require limits on encryption.

303. *Rodriguez v. Google Inc.*, GLOBAL FREEDOM EXPRESSION, https://globalfreedomofexpression.columbia.edu/cases/rodriguez-v-google-inc/ (Spanish language version of opinion available at https://globalfreedomofexpression.columbia.edu/wp-content/uploads/2015/01/544fd356a1da8.pdf).

304. Companies are attempting to create a database of extremist content that violates their terms and to use machine learning to identify content so they can remove it proactively. Dia Kayyali, *European "Terrorist Content" Proposal is Dangerous for Human Rights Globally*, WITNESS BLOG (Dec. 6, 2018), https://blog.witness.org/2018/12/european-terrorist-content-proposal-dangerous-human-rights-globally/.

305. Land, *supra* note 47.

306. Kayyali, *supra* note 304.

online content.[307] This is Danielle Citron's proposal to modify Section 230 to exempt true "bad actors" that have made a business model out of hosting and promoting such speech and often make money from extorting victims before removing harmful content. Imposing liability on a platform for its own conduct would not trigger the human rights concerns identified earlier, since in those instances the platform's incentives are aligned with the objectives of proportionality. The platform is not being asked to regulate the bad conduct of others—which would result in overbroad censorship. Rather, it is being asked to regulate its own bad conduct, to avoid creating platforms that are aimed at and depend for their existence on hosting and promoting harmful speech.

Although 8chan might be an easy case, it will be difficult to draw the line in others. For example, it cannot be that simply engaging in moderation renders a platform culpable in the resulting harm. Gillespie, for example, suggests we might provide immunity to a platform that simply "offers to connect you to friends or followers and deliver what they say to you and what you say to them," but withdraw that immunity "the moment a platform begins to select some content over others, based not on a judgment of relevance to a search query but in the spirit of enhancing the value of the experience and keeping users on the site."[308] This, however, would remove immunity from far too many platforms, including those that are seeking to moderate to meet the expectations and desires of their users.

The hardest cases will be those in which platforms moderate in ways that are designed to promote their profits, but which also in the process magnify the harms of speech. What happens when the business model of a platform contributes to an environment in which harmful speech is profitable? For example, one report contends that YouTube's recommendation algorithm, which "accounts for more than 70 percent of all time spent on the site," promotes the spread of misinformation and propaganda.[309] As the article notes:

> [C]ritics and independent researchers say YouTube has inadvertently created a dangerous on-ramp to extremism by combining two things: a business model that rewards provocative videos with exposure and advertising dollars, and an algorithm that

307. Kim argues in favor of amending Section 230 to provide for baseline liability with immunity safe harbors. Nancy S. Kim, *Website Design and Liability*, 52 JURIMETRICS J. 383, 413-19 (2012). A more conservative approach, however, may be needed because it will be difficult to envision all of the safe harbors needed to protect competing interests.

308. Gillespie, *supra* note 286, at 211.

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

> guides users down personalized paths meant to keep them glued to their screens.[310]

The findings of this study have been contested,[311] and YouTube's changes to its algorithm since the initial study have likely affected the algorithm's impact[312]—changes that are difficult to ascertain since researchers can only study these impacts externally.[313] Nonetheless, the underlying question remains: Should intermediaries lose their immunity if they contribute to the harms of speech in their pursuit of advertising dollars? What happens when they become aware of the effect that their business model is having?

Because a proactive monitoring would have a disproportionate impact on speech, human rights law would favor a duty of care triggered only after notice. Once notified, then, intermediaries would be under an obligation of due diligence—an obligation to investigate the reported content and evaluate the human rights impact of its further dissemination. The level of care required during this human rights impact assessment should ideally be a recklessness standard. Both strict liability and negligence standards would lead to overbroad and disproportionate removals. A recklessness standard, in contrast, would rebut the presumption of liability only if the intermediary knew or should have known of the human rights harms of the speech and failed to take steps to minimize that impact while also protecting the expressive rights of its users. Thus, an intermediary would be liable if it could be shown that it was reckless in its evaluation of the harms of the speech and failed to take steps to minimize that harm—such as by reducing its virality. Finally, because human rights law also requires protection of expression, the intermediary's duty of care should include the obligation to protect the expressive rights of users.[314]

This approach clearly requires the true "bad actor" intermediaries to stop soliciting harmful content and engaging in harmful behavior themselves. But it would also require other intermediaries to change their

---

310. *Id.*

311. Mark Ledwich, *Algorithmic Radicalization—The Making of a New York Times Myth*, MEDIUM (Dec. 27, 2019), https://medium.com/@markoledwich/youtube-radicalization-an-authoritative-saucy-story-28f73953ed17.

312. Anna Zaitsev, *Response to Critique on Our Paper "Algorithmic Extremism: Examining YouTube's Rabbit Hole of Radicalization,"* MEDIUM (Dec. 30, 2019), https://medium.com/@anna.zaitsev/response-to-critique-on-our-paper-algorithmic-extremism-examining-youtubes-rabbit-hole-of-8b53611ce903.

313. Zeynep Tufekci (@zeynep), TWITTER (Dec. 29, 2019, 12:10PM), https://twitter.com/zeynep/status/1211333765051670529.

314. Langvardt, *supra* note 145, at 1376-77. The UK Online Harms White Paper, for example, envisions a regulator that will ensure that platforms respect freedom of expression, although it does not specify a mechanism for doing so. ONLINE HARMS WHITE PAPER, *supra* note 48, at 56.

business models and their algorithms if they realize these models and algorithms are contributing to harm. Although this would limit the extent to which companies can pursue their commercial objectives, public law often requires companies to moderate their pursuit of profit when their actions cause harm. In this way, intermediary regulation could preserve the freedom of expression of their users while tamping down on the virality of speech that magnifies its harm.[315] Thus, this duty of care would be aimed at transforming online hate speech into something that looks more like its offline equivalent—individuals who make hateful comments in small conversations or shout invectives on the street, whose speech quickly fades into oblivion.

2. Specificity and Guidance

Second, any liability imposed on intermediaries must be sufficiently specific about the nature of the speech to be limited and the conditions under which such limitation can occur. As David Kaye notes, limits on discretion in enforcing and applying rules is a critical element of the requirement of legality.[316] Citron similarly argues that failure to adequately define speech can lead to "censorship creep,"[317] extending systems of censorship to include material that is not prohibited by national law.[318] She explains:

> Clarity in the definition, meaning, and application of the terms "hate speech" and "terrorist material" would help contain censorship creep. . . . [U]sers need[] this information to understand their rights and responsibilities when using platforms. Definitional clarity serves another goal: preventing private hate-speech bans from being leveraged to silence legitimate expression.[319]

Tasking an intermediary with a specific and well-defined objective is consistent with human rights law because it reduces the discretion the intermediary might have to substitute its own interests for that of the regulated individual.[320]

---

315. Kim, for example, calls for "structural barriers to speech" that would moderate its harm "without unreasonably restricting it." Kim, *supra* note 97, at 1015.

316. David Kaye, *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/74/486, ¶ 31 (Oct. 9, 2019).

317. Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1050 (2018).

318. *Id.* at 1039 ("Calls to remove hate speech have quickly ballooned to cover expression that does not violate existing European law, including 'online radicalization' and 'fake news.'").

319. *Id.* at 1062.

320. Yildirim v. Turkey, 2012-VI Eur. Ct. H.R. 505, ¶ 64 (noting that prior restraints are only compatible with the European Convention on Human Rights if there is a legal framework "ensuring

Delegations must also ensure that companies are provided with guidance on how laws that are created for offline context should be applied online. As discussed above, online enforcement of hate speech prohibitions is different from offline enforcement in a variety of ways. Further, traditional law enforcement is highly "leaky." Whether someone is prosecuted for hate speech depends on whether they were caught making the statement, whether state officials decide to bring charges, and whether there is enough evidence to obtain a conviction. Social media companies, in contrast, can achieve much more "perfect" enforcement. As a result, the distribution of enforcement is likely to be much different online, with potentially disproportionate effects for human rights. As a result, definitions of hate speech provided to social media platforms will need to be even more precise than their offline equivalents.

3. Accountability Mechanisms

Recent proposals for and critiques of intermediary liability have also focused on accountability mechanisms, such as judicial review, to ensure that intermediary liability does not result in overbroad censorship. The Manila Principles, for example, advocate strongly for shielding intermediaries from liability for third-party content,[321] subject only to orders from judicial authorities to remove content. Under Principle II, intermediaries should only be required to restrict content pursuant to an order issued by an "independent and impartial judicial authority"[322] that complies with certain evidentiary and due process requirements.[323] Principle VI notes that governments "must not use extra-judicial measures to restrict content," including by pressuring changes in terms of service or other voluntary measures.[324]

The Special Rapporteur for Freedom of Expression of the Inter-American Commission on Human Rights has similarly argued that human rights law means that intermediaries should only be required to restrict content "when ordered [to do so] by a court or similar authority that operates with sufficient safeguards for independence, autonomy, and impartiality, and that has the capacity to evaluate the rights at stake and offer the necessary assurances to the user."[325] National courts have also emphasized the importance of judicial review of state removal orders. In

---

both tight control over the scope of bans and effective judicial review to prevent any abuse of power").

321. *Manila Principles*, *supra* note 284 (Principle I(b)). Principle I also recommends that intermediaries never be subject to strict liability or other monitoring obligations. *Id.* (Principle I(c)).

322. *Id.* (Principle II(a)).

323. *Id.* (Principle II-VI).

324. *Id.* (Principle VI(b)).

325. BOTERO MARINO, *supra* note 283, ¶ 106.

2015, for example, the Supreme Court of India "strengthened the safe harbor provisions for Internet intermediaries in section 79 of the IT Act, requiring a court or government order for takedowns under this provision."[326]

Other proposals have envisioned oversight by a non-judicial or administrative authority. For example, the UK White Paper provides that the proposed duty of care would be "overseen and enforced by an independent regulator."[327] This regulator "will have a suite of powers to take effective enforcement action against companies that have breached their statutory duty of care."[328] The White Paper provides that the regulator will set out how companies should fulfill their new legal duty "in codes of practice. If companies want to fulfil this duty in a manner not set out in the codes, they will have to explain and justify to the regulator how their alternative approach will effectively deliver the same or greater level of impact."[329]

Other proposals envision private accountability mechanisms. Special Rapporteur Kaye has proposed "an independent 'social media council', modelled on the press councils that enable industry-wide complaint mechanisms and the promotion of remedies for violations."[330] According to Kaye, "[t]his mechanism could hear complaints from individual users that meet certain criteria and gather public feedback on recurrent content moderation problems such as overcensorship related to a particular subject area."[331] Global Partners Digital has also advocated the creation of an independent oversight board "funded by platforms, but made up of multistakeholder representatives."[332] Facebook has recently created an Oversight Board that will review selected moderation decisions and reverse Facebook's decisions where needed.[333]

The German NetzDG also provides for the creation of private "self-regulation" institutions. These institutions, which must be approved by the German government, would be "funded by several social network providers or establishments, guaranteeing that the appropriate facilities are in place," and "must remain open to the admission of further providers, of

326. *GNI Welcomes Landmark Freedom of Expression Ruling by India*, GLOBAL NETWORK INITIATIVE (Mar. 24, 2015), https://globalnetworkinitiative.org/gni-welcomes-landmark-freedom-of-expression-ruling-by-the-supreme-court-of-india/.

327. ONLINE HARMS WHITE PAPER, *supra* note 48, at 7.

328. *Id.*

329. *Id.*

330. Kaye, *supra* note 81, ¶ 58; *see also* KAYE, *supra* note 40, at 118.

331. Kaye, *supra* note 36, ¶ 58.

332. GLOBAL PARTNERS DIGITAL, A RIGHTS-RESPECTING MODEL OF ONLINE CONTENT REGULATION BY PLATFORMS 6 (2018), https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf.

333. Nick Clegg, *Charting a Course for an Oversight Board for Content Decisions*, FACEBOOK (Jan. 28, 2019), https://newsroom.fb.com/news/2019/01/oversight-board/.

social networks in particular."[334] These self-regulation institutions appear to be "modeled after the German Association for Voluntary Self-Regulation of Digital Media service providers (FSM e.V.), a non-profit association responsible for the protection of minors in the Internet" which handles user complaints for violations of a law protecting minors in the media.[335] Further, intermediaries who comply with decisions of the "self-regulation" body would enjoy safe harbor immunity.[336]

In addition to proposals for specific mechanisms, there have also been widespread calls for greater transparency in order to ensure accountability. Transparency about platform terms of service and their enforcement is needed both to ensure that individuals know what content is allowed and what is prohibited and to enable them to hold the platform accountable for violations. Some of the recent government proposals have incorporated transparency requirements, including both the UK White Paper and NetzDG—although the transparency required in these proposals focuses not on the potential impact on freedom of expression and user rights, but rather on the vigor of platform enforcement.[337]

Finally, any accountability mechanism must be able to deal with the issue of scale. Facebook had 2.5 billion monthly active users as of December 2019.[338] Every minute on Facebook, users post 510,000 comments, update 293,000 statuses, and upload 136,000 photos.[339] Five hundred hours of video are uploaded to YouTube every minute by 2 billion monthly active users.[340] Twitter manages 500 million tweets each day.[341] At this scale, judicial review of each contested piece of content would be impossible. The Manila Principles, for example, recognize the need "to balance this ideal [of judicial review] against the need for

334. NetzDG, *supra* note 54, § 6(5). The analysts at this institution must be independent and possess the necessary expertise, and the institution must have "rules of procedure which regulate the scope and structure of the analysis, stipulate the submission requirements of the affiliated social networks, and provide for the possibility to review decisions." *Id.* §§ 6(1), 6(3)

335. Thomas Wischmeyer, *"What is illegal offline is also illegal online"—The German Network Enforcement Act 2017*, at 9-10 (Sept. 27, 2018), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3256498.

336. *Id.* at 10.

337. NetzDG, *supra* note 54, § 2 (bi-annual reporting on the efforts the platform is making to comply with complaints); ONLINE HARMS WHITE PAPER, *supra* note 48, at 44 (noting that the regulator will have the power to require reporting on "[m]easures and safeguards in place to uphold and protect fundamental rights" but also on the company's terms and conditions, processes for reporting harmful content, the company's use of tools to "identify, flag, block or remove illegal or harmful content" as well as its cooperation with law enforcement).

338. ZEPHORIA DIGITAL MKTG., *supra* note 33.

339. *Id.*

340. Mansoor Iqbal, *YouTube Revenue and Usage Statistics (2019)*, BUS. APPS (Aug. 19, 2019), https://www.businessofapps.com/data/youtube-statistics/; Damien Cave, *Countries Want to Ban "Weaponized" Social Media. What Would That Look Like?*, N.Y. TIMES (Mar. 31, 2019), https://www.nytimes.com/2019/03/31/world/australia/countries-controlling-social-media.html.

341. Cave, *supra* note 340.

expedited action in exceptional circumstances, and also that other legitimate interests that may be impacted by the administrative and financial burden that large quantities of content restriction requests may create."[342] Thus, under the Manila Principles, automated removal or removal pursuant to notice and takedown may be possible, but should be "limited to situations of manifest illegality or where harm to victim is irreparable."[343]

Effective approaches to promoting accountability and transparency are likely to require resort to a variety of mechanisms. The most stringent review should be focused on those areas where the intermediary is exercising delegated state authority. For example, governments might establish quasi-judicial mechanisms to review content moderation that the state is requiring the intermediary to provide. Not every moderation decision would be reviewed; rather, review might be structured like an audit, relying on sampling techniques to generate a group of cases that would allow the reviewing body to evaluate the quality of platform implementation of delegated authority. Legislation should also require platforms to create procedures that allow users to contest adverse platform decisions, and these appeals might also be reviewed. Oversight would thus be targeted rather than comprehensive, but still aimed toward ensuring accountability. Such targeted oversight should be coupled with a legislative framework for and administrative review of platform policies and procedures for engaging in due diligence and mitigation of harm after notice.[344]

4. Moderation by Design

An enabling regulatory environment would also require intermediaries to provide users with more tools to choose the content they want to see. Greater user autonomy has been promoted by David Kaye,[345] Tarleton Gillespie,[346] and Danielle Citron and Ben Wittes,[347] among many others. User controls include tools to mute or block content or users, or to "create closed or private groups, moderated by users themselves."[348] As Kaye

342. MANILA BACKGROUNDER, *supra* note 78, at 24.

343. *Id.* at 17.

344. Langvardt argues that "[a]ny reasonable statutory framework would therefore try to focus judicial, regulatory, and corporate attention on sound content moderation *policies* rather than fussing over individual cases." Langvardt, *supra* note 145, at 1376.

345. Kaye, *supra* note 36, ¶ 60.

346. GILLESPIE, *supra* note 34, at 199-200.

347. Danielle Citron & Benjamin Wittes, *Follow Buddies and Block Buddies: A Simple Proposal to Improve Civility, Control, and Privacy on Twitter*, LAWFARE (Jan. 5, 2017), https://www.lawfareblog.com/follow-buddies-and-block-buddies-simple-proposal-improve-civility-control-and-privacy-twitter.

348. Kaye, *supra* note 36, ¶ 60.

notes, "While content rules in closed groups should be consistent with baseline human rights standards, platforms should encourage such affinity-based groups given their value in protecting opinion, expanding space for vulnerable communities and allowing the testing of controversial or unpopular ideas."[349] Providing users with choices about what they want to hear maximizes both the right to speak and the right to be free from harmful speech—it allows users to say what they want, but gives other users the ability to decide whether they would like to hear it.[350]

Of course, a significant concern with this approach is that it will promote greater isolation by allowing individuals to insulate themselves from opinions and ideas in tension with their own. This may be true, but it would also help users become more aware of the *fact* that they are insulated. Current platform law fosters "bubbles" that are nearly invisible to users because it prioritizes content with which users have interacted in the past. If individuals are required to make a choice—either to be exposed to views consistent with their own, or not—they will be more aware that a choice is being made.

Finally, governments can also promote accountability through design by tackling what Haggart and Tusikov call the elephant in the room—the "systemic conditions that have made commercial online platforms so problematic. Their personalized-advertising, algorithm-fueled, maximized-engagement-at-any-cost business model has played a large role in creating a poisonous online environment."[351] It is a central role of government to rein in companies when their pursuit of profit imposes harm on others. Haggart and Tusikov recommend, among other things, that governments think about tackling these systemic conditions by, for example, "[b]anning personalized advertising, limiting data collection and usage and addressing market-concentration issues."[352] All of these would be significant steps forward in creating a positive regulatory environment for intermediaries.

V. CONCLUSION

Privatizing the regulation of speech comes with enormous consequences that cannot lightly be undone. International human rights law provides a framework for understanding the limits of privatization. State action doctrine ensures that states cannot hide behind private

349. *Id.*

350. Land, *supra* note 221, at 973-77.

351. Blayne Haggart & Natasha Tusikov, *What the U.K.'s Online Harms White Paper Teaches Us About Internet Regulation*, CONVERSATION (Apr. 17, 2019), https://theconversation.com/what-the-u-k-s-online-harms-white-paper-teaches-us-about-internet-regulation-115337.

352. *Id.*

platforms to accomplish their goals, and human rights law prevents broad delegations of authority. This does not mean that delegation is not possible—it just means that the resulting activity must comply with human rights principles, which require limits on authority and safeguards for accountability.

Social media companies are the guardians and curators of our speech. Recognizing both the importance and the potential harms of speech and the need to work cooperatively with platforms, this Article proposes differentiated liability combined with targeted oversight and user-centered design as a model for realistic and balanced regulation of intermediaries. Such limits and safeguards are essential to guard against unaccountable and overzealous censorship not only in places like Germany and the United Kingdom, but also in places like the Philippines, Russia, and Venezuela, which look to regulation in Europe as a model.[353] Human rights law provides a common framework for ensuring that national laws regulating intermediaries strike a balance that is sensitive to local priorities and protects the rights of users while also ensuring robust freedom of expression.

353. *Flawed Social Media Law*, *supra* note 239.

* * *