ARTICLE

FREEDOM OF EXPRESSION IN THE AGE OF ONLINE PLATFORMS:

THE PROMISE AND PITFALLS OF A HUMAN RIGHTS-BASED APPROACH TO CONTENT MODERATION

Barrie Sander[*]

ABSTRACT

In today's digital public sphere, individuals have little choice but to participate on online platforms, whose design choices shape what is possible, content policies influence what is permissible, and personalization algorithms determine what is visible. Ensuring that online content moderation is aligned with the public interest has emerged as one of the most pressing challenges for freedom of expression in the twenty-first century. Taking this challenge as its focus, this Article examines the promise and pitfalls of a human rights-based approach to content moderation—with a specific focus on the choices and challenges that online platforms are likely to confront in adhering to their corporate responsibility to respect human rights in this context. The Article examines three dimensions of a human rights-based approach to platform moderation in particular: a substantive dimension, encompassing the alignment of content moderation rules with international human rights law; a process dimension, encompassing the standards of transparency and oversight that platforms should implement as part of their human

* Postdoctoral Fellow, Fundação Getulio Vargas (FGV), School of International Relations, São Paulo, Brazil, barrie.sander@graduateinstitute.ch. The Author would like to thank Mark Bunting, Daphne Keller, Richard Wingfield, Daragh Murray, Mike Godwin, Evelyn Douek, Thomas Kadri, Molly Land, and Nicolas Suzor, as well as the participants at the 8th Annual Conference of the Cambridge International Law Journal held at the Faculty of Law, University of Cambridge, 20-21 March 2019, for their comments on earlier drafts and presentations of this Article. The Author would also like to acknowledge the funding of Fundação de Amparo à Pesquisa do Estado de São Paulo (FAPESP), which enabled this research to be conducted. All errors remain the Author's own.

rights due diligence processes; and a procedural-remedial dimension, encompassing the procedural guarantees and remediation mechanisms that platforms should integrate within their systems of content moderation. The Article concludes by reflecting on some of the limits of the human rights-based approach and cautioning against viewing human rights as a panacea.

I. INTRODUCTION

The online platform revolution—like the digital revolution from which it emerged—has altered the social conditions of speech. [1] By lowering the cost of generating and sharing information, whilst expanding and diversifying access to both content and conversation, online platforms have created unprecedented possibilities for widespread cultural participation and interaction. At the same time, platforms have also established and enabled new methods of control which serve to both limit and shape what users see and hear on a daily basis.

Until recently, online platforms were inclined to disavow the extent to which they govern speech.[2] Yet, platforms have always been active moderators of online content, with today's largest platforms such as Facebook and YouTube, exerting considerable influence over public discourse around the world. When moderating their sites, online platforms generally perform two functions: [3] first, as content *gatekeepers*, platforms determine which categories of content are allowed and prohibited on their

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

2. *See* Timothy Garton Ash et al., *GLASNOST!: Nine Ways Facebook can Make Itself a Better Forum for Free Speech and Democracy*, REUTERS INST. STUD. J., https://reutersinstitute.politics.ox.ac.uk/sites/default/files/2019-01/Garton_Ash_et_al_Facebook_report_FINAL_0.pdf [https://perma.cc/QL9K-UZ66].

3. GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 18 (2018). *See also* York and Zuckerman, 'Moderating the Public Sphere', in JØRGENSEN (ED), *HUMAN RIGHTS IN THE AGE OF PLATFORMS* (2019) 137, at 140 (referring to "the concepts of *hard control* – a platform's authority over what can be published online – and *soft control* – a platform's authority over what we are likely to see, and what is deprioritized in algorithms that govern a user's views of posts on the network") (emphasis in original).

sites;[4] and second, as content *organizers*, platforms individualize the experiences of their users, highlighting some content over others, through algorithmic personalization.[5]

As the digital public sphere has become increasingly concentrated in the hands of a small number of online platforms, concerns have grown that platform moderation is being driven to a significant extent by corporate imperatives for growth and profit at the expense of the public interest.[6] At the same time, a spate of high-profile controversies, including Russia's cyber influence operation on the 2016 US presidential election and the use of online platforms by members of the Myanmar military as part of the government's campaign of mass violence against the Rohingya,[7] have awakened the public to the potential for platforms to be used to disrupt elections, spread hate and disinformation, and inspire deadly atrocities around the globe.[8]

In this climate, the pertinent challenge has become to identify a way to re-align the private incentives of platform governance with the broader public interest.[9] In a report published in April 2018, the UN Special Rapporteur on Freedom of Expression, David Kaye, charted one possible path forward. Kaye's report set out "a framework for the moderation of user-generated online content that puts human rights at the very centre."[10] According to Kaye, human rights principles can enable online platforms "to create an inclusive environment that accommodates the varied needs and interests of their users while establishing predictable and

4. EMILY B. LAIDLAW, REGULATING SPEECH IN CYBERSPACE: GATEKEEPERS, HUMAN RIGHTS AND CORPORATE RESPONSIBILITY 2 (2015).

5. Zuiderveen Borgesius et al., *Should We Worry About Filter Bubbles?*, 5 INTERNET POL'Y REV., no. 1, 2016, at 2.

6. Rikke Frank Jørgensen, *Human Rights and Private Actors in the Online Domain*, *in* NEW TECHNOLOGIES FOR HUMAN RIGHTS LAW AND PRACTICE 243, 251-53 (2018).

7. *See, e.g.*, YOCHAI BENKLER, ROBERT FARIS, & HAL ROBERTS, NETWORK PROPAGANDA: MANIPULATION, DISINFORMATION AND RADICALIZATION IN AMERICAN POLITICS 235-69 (2018); *see also* Steve Stecklow, *Hatebook: How Facebook is Losing the War on Hate Speech in Myanmar*, REUTERS, (Aug. 15, 2018), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ [https://perma.cc/B65U-F83Y].

8. Mark Bunting, *From Editorial Obligation to Procedural Accountability*: Policy Approaches to Online Content in the Era of Information Intermediaries 3 J. CYBER POL'Y 165, 174 (2018).

9. *Id*.

10. Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, UN Doc A/HRC/38/25, 6 Apr. 2018 [hereinafter Kaye Content Moderation Report], para 2.

consistent baseline standards of behaviour."[11] To this end, Kaye recommends that platforms "should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law […] [and] re-evaluate their content standards accordingly."[12]

Building on the foundations of David Kaye's report, the Article examines the promise and pitfalls of a human rights-based approach to platform moderation—with a specific focus on the choices and challenges that online platforms are likely to confront in adhering to their corporate responsibility to respect human rights in this context. To this end, this Article proceeds in three parts. The Article begins by outlining the current practice of platform moderation, illuminating the mechanics of content moderation, the different influences that shape moderation policies and processes, and the concerns that have been raised about how moderation is conducted by online platforms in practice (Part II). The Article then turns to examine the promise and pitfalls of applying a human rights-based approach to help alleviate such concerns (Part III). The Article examines three dimensions of a human rights-based approach to platform moderation in particular: a *substantive* dimension, encompassing the alignment of content moderation rules with international human rights law; a *process* dimension, encompassing the standards of transparency and oversight that platforms should implement as part of their human rights due diligence processes; and a *procedural-remedial* dimension, encompassing the procedural guarantees and remediation mechanisms that platforms should integrate within their systems of content moderation. The Article concludes by reflecting on some of the limits of the human rights-based approach and cautioning against viewing human rights as a panacea (Part IV).

II. THE PRACTICE OF PLATFORM MODERATION

Over the course of the past decade, the growing indispensability of accessing and participating in online discourse has been paired with an increasing concentration of power and control over the Internet's content layer in the hands of a small

11. *Id*. para. 43.
12. *Id*. para. 70.

number of private online platforms.[13] In today's digital public sphere, individuals have little choice but to participate on these platforms,[14] whose design choices shape what is *possible*, content policies influence what is *permissible*, and personalization algorithms determine what is *visible*.[15] By establishing and enforcing rules of private governance that moderate how ideas and information are exchanged online, today's largest platforms have emerged as "governors of online speech,"[16] "custodians of the public sphere,"[17] and "stewards of public culture."[18]

This Part begins by defining the core characteristics of platforms and explaining the mechanics of content moderation (Section A). The section then turns to identify the different factors that influence how platforms moderate online content (Section B), before revealing some of the concerns that platform moderation has given rise to in practice (Section C).

A. Defining Platforms and Content Moderation

The term *platform* is notoriously vague and ambiguous. To some extent this is because the term tends to vary depending on the context in which it is deployed.[19] The term's elusiveness is also part of its appeal. A wide range of companies have designated themselves as platforms in an attempt to appear neutral and evade regulatory scrutiny.[20] In an effort to elucidate the meaning of the term, the Article relies on the definition recently elaborated by Tarleton Gillespie, who refers to platforms as online sites and services that "host, organize, and circulate users' shared content or social interactions for them," without having produced the bulk of

13. Paul Nemitz, *Constitutional Democracy and Technology in the Age of Artificial Intelligence*, ROYAL SOCIETY PHILOSOPHICAL TRANSACTIONS 1, 2-4 (2018).

14. BRUCE SCHNEIER, DATA AND GOLIATH: THE HIDDEN BATTLES TO CAPTURE YOUR DATA AND CONTROL YOUR WORLD 60-61 (2015).

15. *See generally* NICOLAS P. SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES (2018).

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

17. Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2041 (2018).

18. Tarleton Gillespie, *Platforms Are Not Intermediaries*, 2 GEO. L. TECH. REV. 198, 199 (2018).

19. *See generally* Robert Gorwa, *What is Platform Governance?*, 22 INFO., COMM. & SOC'Y 854 (2019).

20. *See generally* Robyn Caplan, *Content or Context Moderation: Artisanal, Community-Reliant, and Industrial Approaches*, DATA & SOC'Y 8 (Nov. 14, 2018).

that content themselves, built on an infrastructure for processing data for a range of different purposes including the generation of profit, and which "moderate the content and activity of users."[21]

Within the broad parameters of this definition, platforms vary significantly in terms of their functions (e.g., social network sites like Facebook, microblogging providers like Twitter, and video-sharing sites like YouTube), business models (e.g., different types of advertising and subscription-based monetization methods), and size (e.g., in terms of the number and geographical spread of their users and employees).[22] Moreover, platforms are not static, evolving over time to develop new uses, sources of revenue, and communities of users.[23]

Yet, despite their diversity, it is the final element of the definition—content moderation—which constitutes the indispensable and definitional part of what platforms do.[24] In practice, online platforms emerged to simplify the process of navigating the abundance of information available in the digital public sphere. As Gillespie explains, "[t]hough part of the web, […] platforms promise to rise above it, by offering a better experience of all this information and sociality: curated, organized, archived, and moderated."[25] It is for this reason that "moderation is, in many ways, *the* commodity that platforms offer."[26] Without moderation, platforms would be unable to shape user participation into the "right" kind of online experience.[27] "Right," Gillespie observes, "may mean ethical, legal, and healthy, but it also means whatever will promote engagement, increase ad revenue, and facilitate data collection."[28]

---

21. GILLESPIE, *supra* note 3, at 18-21.

22. Caplan, *supra* note 20, at 8-13; *see also* OECD, *An Introduction to Online Platforms and their Role in the Digital Transformation* (OECD, 2019).

23. Sasha Desmaris et al., *Creating a French Framework to Make Social Media Platforms More Accountable: Acting in France with a European Vision* 8 (Mission Report submitted to the French Secretary of State for Digital Affairs, May 2019), https://www.numerique.gouv.fr/uploads/Regulation-of-social-networks_Mission-report_ENG.pdf [https://perma.cc/C9TA-L77A] [hereinafter *French Interim Report*].

24. GILLESPIE, *supra* note 3, at 21.

25. *Id*. at 13.

26. *Id*. (emphasis in original).

27. Gillespie, *supra* note 18, at 202.

28. *Id*.

In practice, platform moderation is operationalized through rules of private governance.[29] Some of these rules are *implicit*, expressed in the code and algorithms that influence the types of social interactions that are possible on a platform, as well as how content is organized, promoted and presented to users—processes that amount to forms of "architectural regulation."[30]

Other rules are *explicit*, such as those documented in public-facing platform community standards and terms of service, as well as non-public internal moderation guidelines.[31] Through these documents, which are subject to continual update and revision, platforms establish restrictions on a range of categories of content—most commonly hate speech, graphic or violent content, sexual content, harassment, copyright, and illegal activity—the interpretation and enforcement of which contribute to a body of "platform law."[32]

Importantly, the explicit rules of platform governance are enforced through recourse to a range of moderation techniques.[33] For certain types of content, moderation occurs prior to publication (*ex ante* moderation). A picture-recognition technology called PhotoDNA, for example, relies upon digital fingerprints (hashes) to automatically detect and prevent the upload of known images of child exploitation.[34] Similarly, Content ID, a technology developed by YouTube, scans uploaded content against a database of content provided by copyright owners, who can decide to block, monetize or track content containing their work.[35]

---

29. Bunting, *supra* note 8, at 172.

30. GILLESPIE, *supra* note 3, at 179 & chapt. 7.

31. *See generally* GILLESPIE, *supra* note 3, chapt. 3; *see also* Klonick, *supra* note 16, at 1630-35.

32. Kaye Content Moderation Report, *supra* note 10, para. 1. *See generally* ARTICLE 19, *Side-stepping rights: Regulating speech by contract* (2018); Molly K. Land, *The Problem of Platform Law: Pluralistic Legal Ordering on Social Media* (2019).

33. *See generally* Klonick, *supra* note 16, at 1635. *See also* GILLESPIE, *supra* note 3, chapt. 4-5.

34. Microsoft, *PhotoDNA*, https://www.microsoft.com/en-us/photodna [https://perma.cc/T8M2-DLWX] (last visited June 25, 2019).

35.Google, *How Content ID Works*, https://support.google.com/youtube/answer/2797370?hl=en [https://perma.cc/2EUF-SGPH] (last visited June 25, 2019).

Moderation also occurs after content has already been published (*ex post* moderation). Relying on a combination of community and automated flagging to detect potentially impermissible content, *ex post* review may be conducted automatically by software and/or manually by human moderators. Facebook, for example, recently confirmed that the platform uses machine learning to assess content that may signal support for the Islamic State of Iraq and Syria ("ISIS") or al-Qaeda, producing a score indicating how likely it is that a post violates the platform's counterterrorism policies.[36] Facebook automatically removes posts where the tool's confidence level indicates that its *decision* will be more accurate than human reviewers. For all other posts, the score system enables Facebook's team of human moderators to prioritize reviewing content that receives the highest scores.

In practice, the precise combination of techniques relied upon as well as the organizational structure of content moderation tend to vary depending on a range of factors including the function, size, resources and policies of the platform. Robyn Caplan, for example, distinguishes three major approaches to platform moderation:[37] *artisanal* approaches, involving smaller-scale case-by-case review of content by teams of between 5 and 200 human moderators (e.g., Medium); *community-reliant* approaches, which typically combine overarching policy decisions by a small team of company employees with a larger group of volunteer human reviewers (e.g., Reddit); and *industrial* approaches, which typically rely upon large-scale bureaucracies of tens of thousands of human reviewers to enforce community standards that are defined by separate policy teams (e.g., Facebook).

Importantly, the techniques and organizational structures of content moderation are not neutral but affect how content is reviewed and which values are prioritized in the process. For example, whereas the hands-on approach of artisanal platforms generally enables greater sensitivity to *context*, the bureaucratized approach of industrial platforms tends to place greater emphasis

36. Monika Bickert & Brian Fishman, *Hard Questions: What Are We Doing to Stay Ahead of Terrorists?*, FACEBOOK NEWSROOM, (Nov. 8, 2018) https://about.fb.com/news/2017/06/how-we-counter-terrorism/ [https://perma.cc/NJ4T-B62P].

37. Caplan, *supra* note 20, at 15-25.

on the value of *consistency*.[38] As the next section reveals, these organizational dynamics constitute just one of a broader range of factors that influence how platform moderation is conducted in practice.

B. Influences Over Platform Moderation

Platform moderation is not a static process but an ongoing negotiation between a plurality of actors,[39] not only the various policy teams and moderators employed or contracted by the platforms themselves, but also user communities, governments, advertisers, mass media organizations, civil society groups, and academic experts. In other words, platforms do not moderate content in a vacuum but are subject to a range of pressures that feed into and shape the substance, processes and procedures of their content moderation policies. In practice, content moderation is driven by at least four sets of influences—*corporate philosophy, regulatory compliance, profit maximization*, and *public outcry*—each of which affects what is possible, permissible, and visible on online platforms.

1. Corporate Philosophy

Platform moderation is at least partially shaped by corporate philosophy.[40] Different platforms aim to provide unique products and services for their users. At the most general level, platforms tailor their architectures to ensure they are suitable for providing particular types of experiences—whether photo and video sharing on Instagram or microblogging on Twitter. At a more granular level, platforms operate more or less permissive speech policies depending on the particular environments they wish to nurture on their sites. These environments are typically grounded in particular corporate values which exert influence over how platforms moderate content on their sites in practice. Facebook, for example, places significant emphasis on "authenticity," a value

38. *Id*.

39. On the characterization of online speech governance as "pluralist," see generally Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 UC DAVIS L. REV., 1149, 1186-93 (2018); Hamilton, *Governing the Global Public Square* (manuscript on file with author).

40. *See also* Klonick, *supra* note 16, at 1625-27.

that informs a number of areas of its content moderation practices including its requirement that people that connect on the platform must use "the name that they go by in everyday life."[41]

Importantly, corporate philosophies tend not to be static but evolve over time. Twitter, for example, initially established a largely hands-off moderation policy, only intervening to moderate content in certain exceptional circumstances. Originally branding itself as "the free speech wing of the free speech party,"[42] Twitter's maximalist approach to freedom of expression ended up creating a toxic hunting ground for minority groups, who found themselves targeted by racists, misogynists and trolls simply for participating on the platform. As one commentator put it, "On Twitter abuse is not just a bug, but—to use the Silicon Valley term of art—a fundamental feature."[43]

Over time, however, Twitter has gradually instituted slightly more restrictive moderation policies on the premise that "freedom of expression means little as our underlying philosophy if we continue to allow voices to be silenced because they are afraid to speak up." [44] In practice, a platform's corporate philosophy is important, not only as a means to develop the platform's user base but also to satisfy the platform's founders and employees, who will typically want to ensure that the company's underlying mission and values are aligned with their own.[45]

---

41. *Community Standards: 17. Misrepresentation*, FACEBOOK, https://www.facebook.com/communitystandards/misrepresentation [https://perma.cc/M8E5-FVG9].

42. Josh Halliday, *Twitter's Tony Wang: 'We Are the Free Speech Wing of the Free Speech Party'*, GUARDIAN (Mar. 22, 2012), https://www.theguardian.com/media/2012/mar/22/twitter-tony-wang-free-speech [https://perma.cc/T6PB-QGFA].

43. Charlie Warzel, "A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment, BUZZFEED NEWS (Aug. 11, 2016), https://www.buzzfeednews.com/article/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s [https://perma.cc/WV27-9SLM].

44. Vijaya Gadde, *Twitter Executive: Here's How we're Trying to Stop Abuse While Preserving Free Speech*, WASH. POST (Apr. 16, 2015), https://www.washingtonpost.com/posteverything/wp/2015/04/16/twitter-executive-heres-how-were-trying-to-stop-abuse-while-preserving-free-speech/ [https://perma.cc/9VJS-7DG8].

45. GILLESPIE, *supra* note 3, at 47.

2. Regulatory Compliance

Beyond corporate philosophy, content moderation is also shaped by the need for platforms to comply with various forms of regulation—whether *mandatory* regulatory measures or more *informal* regulatory pressures.[46] In terms of *mandatory* regulatory measures, states sometimes use their authority to order particular courses of action, for example, platform blocking orders restricting access to particular platforms or content restriction orders requiring the restriction of specific content. States also rely on a combination of content restriction laws and intermediary liability laws to influence the governance of speech on online platforms. Content restriction laws define categories of content that are illegal in particular domestic and regional contexts.[47] In the European Union, for example, illegal content includes incitement to terrorism, xenophobic and racist speech that publicly incites hatred and violence, as well as child sexual abuse.[48] Intermediary liability laws establish the conditions under which platforms may be held liable for illegal content generated by their users. Importantly, the scope and nature of content restriction and intermediary liability laws applicable in any given national context will generally affect how platforms moderate their sites. According to Daphne Keller, for example, experience with intermediary liability laws around the world suggests that legislation which lacks rigorous procedural safeguards, places monitoring obligations on companies to proactively police their platforms, defines the mental state required for liability in broad terms, and/or requires platforms to make context-dependent assessments concerning the legality of complex categories of content such as terrorist recruitment materials or propaganda, will generally result in higher rates of lawful content being removed from platforms through their moderation processes.[49]

---

46. For a useful typology of different types of government action concerning the governance of content on online platforms, see generally Molly K. Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, VIRGINIA J. OF INT'L L. (forthcoming).

47. REBECCA MACKINNON ET AL., FOSTERING FREEDOM ONLINE: THE ROLE OF INTERNET INTERMEDIARIES, 31-36 (2014).

48. European Commission, Tackling Illegal Content Online: Towards an Enhanced Responsibility of Online Platforms, at 2, COM (2017) 555 final (Sept. 28, 2017).

49. Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money*, HOOVER INSTITUTION, Oct. 31, 2018, at 18-20.

In addition to mandatory regulatory measures, various forms of *informal regulatory pressure* have also been exerted over the content moderation practices of platforms. Three forms of pressure have proven particularly prevalent in practice. First, special units have been established in certain jurisdictions to flag potentially illegal content to platforms for their voluntary evaluation against their terms of service. Europol, for example, has established an Internet Referral Unit ("IRU") with a mandate to refer terrorist and violent extremist content to online service providers for their voluntary review. Between July 2015 and December 2017, the IRU made 44,807 decisions for referral of terrorist content, with a removal success rate of ninety-two percent.[50]

Second, regulatory institutions have sometimes reached informal agreements with online platforms to establish particular standards in their content moderation practices and meet specific targets. For example, pursuant to the 2016 *Code of Conduct on Countering Illegal Hate Speech Online* agreed between the European Commission and Facebook, Microsoft, Twitter, and YouTube, participating companies committed to collaborate with "trusted reporters"—particularly civil society groups but in practice also specialized law enforcement departments who will notify platforms of the existence of illegal hate speech—review the majority of valid removal notifications in less than twenty-four hours, and remove or disable access to such content if found to violate their terms of service or national laws.[51] In other instances, voluntary agreements have been reached with respect to content that is lawful but nonetheless deemed harmful or undesirable.[52] In 2018, for example, the European Commission unveiled a *Code of Practice on Online Disinformation*, which establishes self-regulatory standards for social media platforms and the advertising industry to fight disinformation worldwide.[53]

Finally, content moderation practices have also been shaped by various forms of jawboning through public appeals for

50. EU Internet Referral Unit, *Transparency Report 2017*, at 5 (Sept 12, 2018).

51. European Commission Press Release IP/16/1936, European Commission and IT Companies Announce Code of Conduct on Illegal Online Hate Speech, (May 31, 2016).

52. Raso et al., *Artificial Intelligence & Human Rights: Opportunities & Risks*, BERKMAN KLEIN CENTER FOR INTERNET & SOCIETY, Sept. 25, 2018, at 37-38.

53. EU Code of Practice on Disinformation (2018).

platforms to alter their moderation processes and practices in particular ways or face the prospect of future regulation. Danielle Citron, for example, has examined the impact of "the shadow of threatened regulation" by the EU on the content moderation policies of online platforms.[54] In particular, the adoption by Facebook, Microsoft, Twitter and YouTube of a shared industry database of hashes for terrorist content appears to have been timed to diminish the prospect of future regulation that was feared might follow the European Commission's critical review of their compliance with the *Code of Conduct on Countering Illegal Hate Speech Online*.[55]

As this analysis indicates, a range of regulatory tools have been relied upon to incentivize online platforms to put in place systems of content moderation that meet particular requirements, standards, and targets. It is important to recognize, however, that not all governments have sufficient leverage to secure the attention and successfully ensure compliance of platforms with their regulatory demands—typically only those that control access to the most commercially valuable markets.[56]

3. Profit Maximization

Since the digital public sphere is predominantly controlled by private platforms, content moderation is also influenced to a significant extent by the corporate imperative to maximize profits. Notably, many of today's largest online platforms rely on a business model that involves the sale of human attention. As the cost of creating and distributing content has radically declined and the speed at which information can be disseminated has become

54. Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1045-49 (2018). *See also* Balkin, *supra* note 39, at 1177 (referring to the State practice of "jawboning – urging digital infrastructure operators to do the right thing and block, hinder, or take down content"); Robert Gorwa, *The Platform Governance Triangle: Conceptualizing the Informal Regulation of Online Conten*t, 8 DATA & SOCIETY (2019).

55. Citron, *supra* note 54, at 1048. On the concerns raised by opaque and unaccountable forms of collaboration between online platforms, see generally Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMENDMENT INSTITUTE (2020).

56. *See* Andrew Keane Woods, *Litigating Data Sovereignty*, 128 YALE L. J. 328, 405 (2018); Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, *Aegis Series Paper No. 1902*, at 7 (2019); Hamilton, *supra* note 39.

supercharged,[57] informational scarcity within the mass media public sphere has been superseded by what Tim Wu terms "attentional scarcity" within the digital public sphere.[58] Drawing on their enormous social networks, major social media platforms such as Facebook and YouTube generate revenue and profits by monetizing the attention of their users.[59]

Through a phenomenon that Shoshana Zuboff calls "surveillance capitalism,"[60] many online platforms enable people to connect and communicate with each other around the world in exchange for surveilling their online expressions and behavior.[61] In this context, surveillance is driven by a financial imperative: platforms collect large swathes of data in order to monetize it by selling targeted advertising.[62] Using big data analytics, platforms develop highly specific and detailed digital dossiers about their users so that advertising can be narrowly tailored according to their demographics and inferred interests.[63] Data collection also enables platforms to curate and present content to their users in ways that aim to improve engagement with their sites.[64] Increasing user engagement is financially lucrative for online platforms: as users spend more time and attention on their sites, platforms can collect ever more behavioral data, improve their targeted advertising and engagement capabilities, and grow their advertising revenue.[65]

---

57. *See* Claire Wardle & Hossein Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking,* at 11-12 (2017).

58 . Tim Wu, *Is the First Amendment Obsolete?, in Emerging Threats* (2017) https://knightcolumbia.org/content/tim-wu-first-amendment-obsolete [https://perma.cc/522Q-9T9D].

59. *Id*.

60. *See generally* Shoshana Zuboff, *Big Other: Surveillance Capitalism and the Prospects of an Information Civilization*, 30 J. INFO. TECH. 75 (2015)

61. Jack Balkin, *Fixing Social Media's Grand Bargain*, *Aegis Series Paper No. 1814* at 3 (2018) https://www.hoover.org/sites/default/files/research/docs/balkin_webreadypdf.pdf [https://perma.cc/2EML-NUPR].

62. Nathalie Marechal, *Targeted Advertising Is Ruining the Internet and Breaking the World*, VICE: MOTHERBOARD, (Nov. 16, 2018), https://www.vice.com/en_us/article/xwjden/targeted-advertising-is-ruining-the-internet-and-breaking-the-world [https://perma.cc/3X8X-59L9].

63. DIPAYAN GHOSH & BEN SCOTT, #DIGITALDECEIT: THE TECHNOLOGIES BEHIND PRECISION PROPAGANDA ON THE INTERNET, 5-12 (2018).

64. JAMIE BARTLETT, THE PEOPLE VS TECH: HOW THE INTERNET IS KILLING DEMOCRACY (AND HOW WE CAN SAVE IT), 25-26 (2018).

65. Balkin, *supra* note 61, at 3.

In terms of content moderation, surveillance capitalism incentivizes online platforms to moderate content in ways that aim to maximize both user engagement and advertising revenue on their platforms. In this vein, changes to content moderation policies have sometimes been driven by the demands of the advertising industry. In 2013, for example, Facebook updated its policy on sexually violent content after fifteen advertisers pulled their advertising in response to images that glorified rape and domestic violence appearing on the platform.[66] In addition, content moderation has also been guided by the goal of increasing the engagement of users. This is particularly evident in the algorithmic personalization that many online platforms offer in an effort to keep users glued to their sites—a product of which has been the promotion of emotionally charged, extreme and inflammatory content.[67] Although profit is not the sole driver of content moderation on online platforms, over the years, it has become readily apparent that content moderation has become tied to a business model that incentivizes the maximization of user engagement, data surveillance, and targeted advertising for the generation of revenue and growth.

4. Public Outcry

Platform moderation can also be influenced by public outcry. In particular, when collective action by users, civil society groups, and/or members of the general public concerning particular moderation policies has been paired with significant media coverage and/or litigation, platforms have sometimes—though by no means always—been responsive to the concerns raised.[68] A high-profile example that illustrates both the potential and limits of public collective action concerns the evolution of Facebook's moderation policy regarding breastfeeding photos. In 2008, Facebook became the target of an 80,000-plus protest by

66. Laura Stampler, *Facebook Will Block Photos Celebrating Rape Following Ad Boycott*, BUSINESS INSIDER (May 28, 2013), https://www.businessinsider.com/facebook-fbrape-ad-boycott-2013-5 [https://perma.cc/A8KH-68G8].

67. *See, e.g.*, Zeynep Tufekci, *YouTube, the Great Radicalizer*, N.Y. TIMES: OPINION, (Mar. 10, 2018) https://www.nytimes.com/2018/03/10/opinion/sunday/youtube-politics-radical.html [https://perma.cc/FU4E-ZRZT]; Ronald Deibert, *Three Painful Truths About Social Media*, 30 J. DEMOCRACY 25, 31-34 (2019).

68. Klonick, *supra* note 16, at 1652-55.

angered mothers after breastfeeding photos were removed from its platform.[69] Initially, Facebook merely attempted to clarify rather than alter its policy. In 2009, the platform explained that it only intervened when a photo contained a fully exposed breast, citing concerns about allowing such photos on a site where the minimum age is 13-years-old.[70] In 2012, a second wave of protests and the leak of Facebook's internal moderation guidelines garnered further press coverage around the issue.[71] However, protesters would ultimately have to wait until 2015 before Facebook finally yielded to pressure and updated its moderation rules, explaining that the platform would restrict some images of female breasts if they included the nipple but would "always allow photos of women actively engaged in breastfeeding or showing breasts with post-mastectomy scarring."[72] The struggle against Facebook's moderation policy on breastfeeding photos demonstrates how collective action can sometimes provoke changes in the rules of social media platforms. At the same time, the episode also reveals that significant media coverage and protests spanning a number of years may be required before a platform is willing to implement even slight alterations to its policies.

C. Concerns Over Platform Moderation

Although subject to a range of external pressures, major online platforms such as Facebook and YouTube wield enormous influence over the digital public sphere, acting as a gateway for information and expression around the world. Initially emerging as a means to tame the disorder of the open web, today's largest online platforms have gradually developed increasingly intricate systems of moderation that influence what is possible, permissible, and visible online. In practice, however, platform moderation

69. Mark Sweney, *Mums Furious as Facebook Removes Breastfeeding Photos*, GUARDIAN (Dec. 30, 2008), https://www.theguardian.com/media/2008/dec/30/facebook-breastfeeding-ban [https://perma.cc/2Y2B-9HDA].

70. GILLESPIE, *supra* note 3, at 160.

71. *Id*. at 162*ff.*

72. Vindu Goel, *Facebook Clarifies Rules on What It Bans and Why*, N.Y. TIMES: BITS, (Mar. 16, 2015), https://bits.blogs.nytimes.com/2015/03/16/facebook-explains-what-it-bans-and-why/ [https://perma.cc/4W2W-R4KF].

policies have given rise to a range of *substantive*, *process*, and *procedural-remedial* concerns.[73]

1. Substantive Concerns

In terms of substantive concerns, online platforms have often found themselves on the receiving end of criticism for adopting content standards that are deemed either too restrictive or too permissive. Given the size and diversity of the communities on today's leading online platforms, criticism of content permissibility standards is to some extent inevitable. At the same time, platforms have often appeared inattentive or indifferent to the trade-offs involved in the policies they adopt.

Facebook, for example, has been criticized for failing to carefully consider the implications of the platform's authentic name requirement. Ostensibly aimed at protecting users from online harassment, the policy—which initially required users to use their legal names on the platform—has proven culturally biased and hazardous for groups as diverse as drag queens, human rights activists, victims of crime, and minority groups that rely on pseudonyms to protect themselves from physical harm and danger.[74] In addition, when combined with Facebook's community flagging system, the policy has also facilitated "organised reporting sprees" against political activists and other vulnerable groups—effectively enabling forms of abuse that the policy was designed to deter.[75]

Beyond the relative restrictiveness of content standards, concerns have also arisen that the imprecision and ambiguity of many moderation rules can render them vulnerable to censorship

73. *See generally* GILLESPIE, *supra* note 3; ZEYNEP TUFEKCI, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST 132-64 (2017); SIVA VAIDHYANATHAN, ANTI-SOCIAL MEDIA: HOW FACEBOOK DISCONNECTS US AND UNDERMINES DEMOCRACY (2018); MIKE GODWIN, THE SPLINTERS OF OUR DISCONTENT: HOW TO FIX SOCIAL MEDIA AND DEMOCRACY WITHOUT BREAKING THEM (2019); DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET (2019).

74. Kaye Content Moderation Report, *supra* note 10, ¶ 30; GILLESPIE, *supra* note 3, at 62-63; ARTICLE 19, *supra* note 32, at 30.

75. Dia Kayyali, *Facebook's Name Policy Strikes Again, This Time at Native Americans*, ELECTRONIC FRONTIER FOUND. (Feb. 13, 2015), https://www.eff.org/deeplinks/2015/02/facebooks-name-policy-strikes-again-time-native-americans [https://perma.cc/9SYL-REDJ].

creep, inconsistent application, and discriminatory enforcement.[76] David Kaye, for example, has recently observed that the vagueness of platform hate speech and harassment policies "has triggered complaints of inconsistent policy enforcement that penalizes minorities while reinforcing the status of dominant or powerful groups."[77] Inconsistent application of moderation rules may also be exacerbated by the time and resource constraints placed on human moderators, as well as their lack of knowledge about the specific linguistic and cultural contexts in which content is shared.[78]

Inadequate sensitivity to local contexts in the formulation, application and enforcement of moderation rules can leave platforms open to instrumentalization for the spread of hate speech and disinformation, with the attendant risk of triggering or fueling offline discrimination and violence.[79] The Independent International Fact-Finding Mission on Myanmar, for example, recently concluded that death threats, incitement to violence and discrimination, and online harassment against the Rohingya minority group had become "common features" on social media platforms in Myanmar, with Facebook, in particular, emerging as "a useful instrument for those seeking to spread hate in a context where for most users Facebook *is* the Internet."[80]

---

76. *See* Citron, *supra* note 54, at 1051; Kaye Content Moderation Report, *supra* note 10, para. 26.

77. Kaye Content Moderation Report, *supra* note 10, para. 27; Keller, *supra* note 49, at 24; ARTICLE 19, *supra* note 32, at 16.

78. *See* Kaye Content Moderation Report, *supra* note 10, at 11.

79. *See generally* Molly K. Land & Rebecca J. Hamilton, *Beyond Takedown: Expanding the Tool Kit for Responding to Online Hate*, *in* PROPAGANDA AND INTERNATIONAL CRIMINAL LAW: FROM COGNITION TO CRIMINALITY 143 (2019); AM. BAR ASS'N CTR. FOR HUMAN RIGHTS, INVISIBLE THREATS: MITIGATING THE RISK OF VIOLENCE FROM ONLINE HATE SPEECH AGAINST HUMAN RIGHTS DEFENDERS IN GUATEMALA (2019), https://www.americanbar.org/content/dam/aba/administrative/human_rights/invisible-threats-guatemala-may-2019.pdf [https://perma.cc/99B5-LG9B]; Evelyn Douek, *Why Were Members of Congress Asking Mark Zuckerberg About Myanmar? A Primer.*, LAWFARE (Apr. 26, 2018, 7:00 AM), https://www.lawfareblog.com/why-were-members-congress-asking-mark-zuckerberg-about-myanmar-primer [https://perma.cc/JQP5-BF5F]; Dia Kayyali, *Alex Jones, Myanmar, and Free Expression Online*, WITNESS BLOG (Sept. 11, 2018), https://blog.witness.org/2018/09/alex-jones-myanmar-online-free-expression/ [https://perma.cc/9PJH-YX49].

80. Rep. of the Detailed Findings of the Indep. Int'l Fact-Finding Mission on Myanmar, U.N. Doc A/HRC/39/CRP.2, at 339-43 (Sept. 17, 2018); Rep. of the Indep. Int'l Fact-Finding Mission on Myanmar, U.N. Doc A/HRC/39/64, at 14 (Sept. 12, 2018). For further discussion of the responsibilities of online platforms in mass atrocity contexts, see

In recent years, automation and algorithmic technologies have increasingly been touted as possible solutions to the challenges of inconsistent application of content moderation rules.[81] At present, however, such tools remain no substitute for human judgment, particularly where detailed assessments of context are required.[82] As Daphne Keller explains, "no reputable experts suggest that filters are good enough to be put in charge of deciding what is illegal in the first place," their function currently limited to identifying "duplicates of specific material that a human previously flagged." [83] And even then, the risk remains that algorithmic decision-making may be grounded in datasets that are based on discriminatory assumptions, generating inbuilt biases that are difficult to detect and risk marginalizing and disproportionately targeting minority groups. [84] Importantly, moderation biases are not without consequence, potentially triggering feelings of alienation, frustration, and moral outrage within individuals and communities whose content is erroneously removed or restricted.[85]

---

generally Jenny Domino, *Crime as Cognitive Constraint: Facebook's Role in Myanmar's Incitement Landscape and the Promise of International Tort Liability*, CASE WESTERN RESERVE J. INT'L L. (*forthcoming* 2020); Shannon Raj Singh, *Move Fast and Break Societies: the Weaponisation of Social Media and Options for Accountability Under International Criminal Law*, 8 CAMBRIDGE INT'L L. J. 331. (2019).

81. *See, e.g.*, *Tackling Illegal Content Online*, *supra* note 48, at 12-13, 19 (encouraging the use of automatic detection and filtering technologies).

82. Keller, *supra* note 49, at 5-8. On the concerns raised by the use of automation in content moderation, see generally Hannah Bloch-Wehba, *Automatic in Moderation,* CORNELL INT'L L. J. (*forthcoming* 2020); Robert Gorwa, Reuben Binns, & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOCIETY (2020); Emma Llansó, Joris van Hoboken, Paddy Leerssen & Jaron Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression*, TRANSATLANTIC WORKING GROUP, 26 Feb. 2020.

83. Keller, *supra* note 49, at 7.

84. *See, e.g.*, David Kaye, Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, U.N. Doc A/73/348, at 8-9 29 Aug. 2018 [hereinafter *Kaye AI Report*]; COUNCIL OF EUR., COMM. OF EXPERTS ON INTERNET INTERMEDIARIES, ALGORITHMS AND HUMAN RIGHTS: STUDY ON THE HUMAN RIGHTS DIMENSIONS OF AUTOMATED DATA PROCESSING TECHNIQUES (IN PARTICULAR ALGORITHMS) AND POSSIBLE REGULATORY IMPLICATIONS 26-28 (2018), https://edoc.coe.int/en/internet/7589-algorithms-and-human-rights-study-on-the-human-rights-dimensions-of-automated-data-processing-techniques-and-possible-regulatory-implications.html [https://perma.cc/5KHZ-67E9] [hereinafter CoE Report].

85. *See generally* Citron, *supra* note 54, at 1058-61; Keller, *supra* note 49, at 22-24.

2. Process Concerns

Beyond substantive concerns, a number of process concerns have also been raised, centered on issues relating to platform transparency and oversight, as well as the engagement of platforms with different stakeholders. In terms of *platform transparency and oversight*, ever since Google released its first transparency report in 2010, the number of companies producing such reports has increased year-on-year.[86] To date, however, the level of detail contained in transparency reports has proven both variable and inadequate. Ranking Digital Rights' *2018 Corporate Accountability Index*, for example, found serious deficiencies in the quality of information disclosed by companies with respect to the volume and nature of content and accounts removed or restricted for violating platform terms of service, the processes used by platforms to identify violations including whether priority consideration is given to flagging by governments or private individuals, and the number and nature of government and private requests to restrict content or accounts received via formal or official channels.[87]

Inadequacies in platform transparency reports are illustrative of what Sarah Roberts has referred to as a "logic of opacity" that pervades content moderation processes.[88] For instance, platforms have typically been reluctant to reveal details about the human workforce that undertakes platform moderation, including the nature of their work, the stressful conditions under which they review content, and significant shortcomings in the support they receive.[89] The logic of opacity also extends to algorithmic decision-making within platform moderation processes. Dia Kayyali, for example, argues that platforms have generally fallen short of

86. *See generally Transparency Reporting Index*, ACCESS NOW, https://www.accessnow.org/transparency-reporting-index/ [https://perma.cc/XR5C-JR6Z] (last visited Feb. 13, 2020).

87. RANKING DIG. RIGHTS, 2018 CORPORATE ACCOUNTABILITY INDEX 51-60 (2018), https://rankingdigitalrights.org/index2018/assets/static/download/RDRindex2018report.pdf [https://perma.cc/29QM-BRES].

88. Sarah T. Roberts, *Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation*, FIRST MONDAY (Mar. 5, 2018), https://journals.uic.edu/ojs/index.php/fm/article/view/8283/6649 [https://perma.cc/WX46-A6EL].

89. *See generally* SARAH T. ROBERTS, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

providing "the most basic assurances of algorithmic accountability or transparency, such as accuracy, explainability, fairness, and auditability."[90]

Transparency concerns have also been raised with respect to platform advertising practices. Platforms market their advertising services as enabling advertisers to reach not only larger but also more targeted audiences according to a range of demographic and inferred characteristics.[91] However, controversies surrounding the 2016 US presidential election and the UK Brexit referendum, for example, have brought to light the potential for these services to be repurposed by political actors in an effort to tailor and target messages at narrow categories of prospective voters in ways that may be corrosive to democracy.[92]

In addition to inadequate transparency and oversight, concerns have also arisen over the processes that social media companies have established to manage *stakeholder engagement*. For example, online platforms have been criticized for adopting a piecemeal approach to resisting repressive regulatory arrangements with States. Of particular concern are informal forms of cooperation that have been established between platforms and law enforcement agencies, which encourage platforms to respond to removal requests within narrow time-frames by evaluating compliance with their terms of service. Such schemes not only incentivize platforms to sacrifice thoughtful deliberation in favor of speed but also circumvent the rule of law by enabling States to avoid seeking the removal of illegal content through formal legal avenues such as domestic courts.[93] In addition, where such arrangements are based on inadequately defined terms such as "hate speech," they may serve as pretexts for

90. 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/ [https://perma.cc/2SKQ-T9NC]; *see also* Lorna McGregor, Daragh Murray & Vivian Ng, *International Human Rights Law as a Framework for Algorithmic Accountability*, 68 INT'L & COMP. L.Q. 309, 317-20 (2019).

91. For a skeptical view, see BENKLER, FARIS, & ROBERTS, *supra* note 7, at 276-79.

92. *See generally* Vaidhyanathan, *supra* note 73, at 146-74.

93. ARTICLE 19, *supra* note 32, at 16-17; Lucie Krahulcova, *Europol's Internet Referral Unit Risks Harming Rights and Feeding Extremism*, ACCESS NOW (June 17, 2016, 6:11 AM), https://www.accessnow.org/europols-internet-referral-unit-risks-harming-rights-isolating-extremists/ [https://perma.cc/9T8R-Y75V].

governments to request platforms to suppress legitimate debate.[94] And by leveraging platforms' own terms of service—which are typically drafted to apply globally—as proxies for illegality, States can also use these types of arrangements to ensure that content is removed worldwide rather than merely on a country-by-country basis.[95]

Beyond concerns over their relationships with States, platforms have also failed to put in place structured systems for engaging with their users and other relevant stakeholders. As a result, the responsiveness of platforms to concerns raised about their content moderation policies has been somewhat inconsistent. Twitter, for example, has relied on the subjective notion of "newsworthiness" to explain why many of US President Donald Trump's most controversial tweets have not been removed from its platform despite seemingly contravening its terms of service.[96] Meanwhile, Facebook only saw fit to re-instate a famous and historical photo of a nine-year-old Vietnamese girl running naked following a napalm attack once the editor and CEO of a major Norwegian newspaper wrote a letter in protest.[97] These examples suggest that platforms may be susceptible to giving preferential treatment in their content moderation practices to those with public influence or financial power.[98]

---

94. FREEDOM HOUSE, FREEDOM ON THE NET: MANIPULATING SOCIAL MEDIA TO UNDERMINE DEMOCRACY 12-15 (Nov. 2017), https://freedomhouse.org/report/freedom-net/freedom-net-2017 [https://perma.cc/68DF-HM64].

95. Hannah Bloch-Wehba, *Terrorist Speech and Global Platform Governance*, BALKINIZATION (Aug. 22, 2018), https://balkin.blogspot.com/2018/08/terrorist-speech-and-global-platform.html [https://perma.cc/R8WU-693U]; *see also* Citron, *supra* note 54, at 1056-57 (describing how the shared industry database of violent terrorist content "has the potential to blacklist content across the world"); Douek, *supra* note 55, at 23-31 (outlining how informal and opaque forms of collaboration between platforms may exacerbate existing concerns with content moderation practices by compounding accountability deficits, creating a false patina of legitimacy for their decisions, augmenting the power of the largest platforms, and suggesting a false consensus on where lines should be drawn regarding online speech governance).

96. Abby Ohlheiser, *The 3 Loopholes That Keep Trump's Tweets on Twitter*, WASH. POST (July 23, 2018), https://www.washingtonpost.com/news/the-intersect/wp/2018/01/03/the-3-loopholes-that-keep-donald-trumps-tweets-on-twitter/ [https://perma.cc/QBU4-KPY4].

97. Klonick, *supra* note 16, at 1654-55.

98. ARTICLE 19, SELF-REGULATION AND "HATE SPEECH" ON SOCIAL MEDIA 16 (2018), https://www.article19.org/wp-content/uploads/2018/03/Self-regulation-and-%E2%80%98hate-speech%E2%80%99-on-social-media-platforms_March2018.pdf [https://perma.cc/S69D-PHTW]; Klonick, *supra* note 16, at 1654- 55, 1665-66.

Failures to adequately consult relevant stakeholders over platform moderation policies can also generate avoidable errors. In 2017, for example, YouTube's adoption of a new machine-learning algorithm designed to monitor extremist content resulted in the platform removing hundreds of thousands of channels and videos documenting the civil war in Syria, including content that is of potentially significant value to human rights investigators.[99] While the platform subsequently worked closely with human rights groups to restore some of this content, its *ad hoc* restoration process was itself criticized for favoring groups and individuals in Europe and the United States with closer ties to the platform.[100] Both the erroneous removal of content and biases in the restoration process could arguably have been avoided or at the very least mitigated had YouTube consulted more widely with relevant civil society groups in advance before changing its moderation processes.

3. Procedural-Remedial Concerns

Finally, platform moderation policies have also proven procedurally and remedially deficient in a number of respects. Concerns include inadequate user notification that content has been removed or flagged, or an account penalized, insufficient notification of the reasons for such actions, limited appeals processes, and untimely and insufficient remedies for wrongful removals.[101] These issues are all the more pressing given that platform moderation effectively amounts to a new privatized and digital form of prior restraint over public speech. As Jack Balkin explains, platform practices of blocking and removal generally occur "without any judicial determination of whether their speech is protected or unprotected, without any Bill of Rights protections,

99. Dia Kayyali & Raja Althaibani, *Vital Human Rights Evidence in Syria is Disappearing from YouTube*, WITNESS (Aug. 30, 2017), https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/ [https://perma.cc/2D2F-Q64J].

100. Avi Asher-Schapiro, *YouTube and Facebook are Removing Evidence of Atrocities, Jeopardizing Cases Against War Criminals*, INTERCEPT (Nov. 2, 2017), https://theintercept.com/2017/11/02/war-crimes-youtube-facebook-syria-rohingya/ [https://perma.cc/3UE2-JM2J].

101. JAMILA VENTURINI ET AL., TERMS OF SERVICE AND HUMAN RIGHTS: AN ANALYSIS OF ONLINE PLATFORM CONTRACTS 58 (2016); *See generally How to Appeal,* ONLINE CENSORSHIP, https://onlinecensorship.org/resources/how-to-appeal [https://perma.cc/SXQ3-778P] (last visited Feb. 14, 2020) (for a useful overview of platform appeals processes).

without any due process rights to a hearing before the action is taken, or indeed, without any obligation to consider and resolve end-user objections promptly."[102]

III. THE PROMISE AND PITFALLS OF A HUMAN RIGHTS-BASED APPROACH TO PLATFORM MODERATION

As anxieties over platform moderation have risen sharply in recent years, online platforms have begun to open up about the substance, processes, and procedures of their content moderation policies and to acknowledge the importance of taking an expanded view of their responsibilities.[103] Accompanying this shift in tone, today's largest platforms have also begun to hint at the influence of human rights law within their moderation processes. [104] Facebook's Vice-President of Policy Solutions, for example, recently confirmed that the platform's moderation teams already "look for guidance in documents like Article 19 of the International Covenant on Civil and Political Rights" ("ICCPR") in determining where to draw the line on freedom of expression with respect to user content. [105] The company also revealed that it "look[s] to international human rights standards" to determine whether the content that would otherwise violate the platform's community standards should be allowed because it is newsworthy and in the public interest.[106] Beyond Facebook, Twitter CEO Jack Dorsey has

102. Balkin, *supra* note 17, at 2018.

103. *See, e.g.*, Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK NEWSROOM (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/YN8R-SQWE].

104. *See generally* Rikke Frank Jørgensen, *Rights Talk: In the Kingdom of Online Giants*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 163 (2019) (on the different narratives related to online platforms' commitment to respect human rights).

105. Richard Allen, *Hard Questions: Where Do We Draw the Line on Freedom of Expression*, FACEBOOK NEWSROOM (Aug. 9, 2018), https://about.fb.com/news/2018/08/hard-questions-free-expression/ [https://perma.cc/RWX2-T99T].

106. Monika Bickert, *Updating the Values That Inform Our Community Standards*, FACEBOOK NEWSROOM (Sept. 12, 2019), https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/ [https://perma.cc/P6AA-NHEK]; *see generally* E. Douek, *Why Facebook's 'Values' Update Matters*, LAWFARE (Sept. 16, 2019), https://www.lawfareblog.com/why-facebooks-values-update-matters [https://perma.cc/A6QG-3FZE].

also acknowledged that his company's values should be rooted in human rights law.[107]

Whether these views signal the beginnings of a more general turn towards a human rights-based approach to content moderation remains to be seen. In any case, momentum is clearly building in support of such an approach.[108] Beyond David Kaye's landmark report mentioned in the introduction to the Article, a human rights-based approach to platform moderation has received the support of a growing number of civil society actors,[109] whilst references to human rights have also begun to appear in a number of statements and regulatory initiatives concerning online speech governance supported by States.[110]

Against this background, this section begins by defining the contours of a human rights-based approach and explaining its value and limitations in the platform moderation context (Section A). The section then turns to elaborate some of the choices and challenges that platforms are likely to confront in operationalizing a human rights-based approach, focusing in particular on the substantive (Section B), process-related (Section C), and procedural-remedial (Section D) dimensions of content moderation. The aim is to illuminate both the promise and the pitfalls of a human rights-based approach to platform moderation.

---

107. Jack Dorsey TWITTER (August 10, 2018), accessible here [https://perma.cc/A297-PPMA].

108. *See generally* Dennis Redeker, Lex Gill, & Urs Gasser, *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*, 80 INT'L COMMUNICATION GAZETTE 302 (2018); NICOLAS P. SUZOR, *LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES* (Submitted Version 2019).

109. *See, e.g.*, Jillian C. York & Corynne McSherry, *Content Moderation is Broken. Let Us Count the Ways,* ELECTRONIC FRONTIER FOUNDATION DEEPLINKS BLOG (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways [https://perma.cc/YAN5-TBZ5]; ACCESSNOW, *Protecting Free Expression in the Era of Online Content Moderation* (May 2019), https://www.accessnow.org/cms/assets/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf [https://perma.cc/ZEH2-T5T7]; ARTICLE 19, *supra* note 32.

110. *See, e.g.*, CHRISTCHURCH CALL TO ELIMINATE TERRORIST & VIOLENT EXTREMIST CONTENT ONLINE, https://www.christchurchcall.com/ [https://perma.cc/UE4Q-X9YT] (last visited Feb. 14, 202) ("All action on this issue must be consistent with principles of a free, open and secure internet, without compromising human rights and fundamental freedoms, including freedom of expression"); and French Interim Report, *supra* note 23, at 19 ("the objectives of the regulatory system must be to defend the exercise of all rights and freedoms on social media platforms").

A. Defining a Human Rights-Based Approach to Platform Moderation

The starting point for defining a human rights-based approach to platform moderation is the *United Nations' Guiding Principles on Business and Human Rights* ("UNGP").[111] The UNGP elaborates a three-pillar framework, generally referred to as "Protect, Respect and Remedy," whose purpose is to prevent, mitigate and redress business-related human rights abuses.[112] The *first pillar* outlines the State duty to *protect* against human rights abuses within their territory and/or jurisdiction by third parties, including business enterprises. The *second pillar* elaborates the corporate responsibility to 'respect' human rights by avoiding infringing on the human rights of others and addressing adverse human rights impacts with which they are involved. The *third pillar* addresses the responsibilities of States and businesses to ensure victims have adequate access to remedies.[113]

Although the UNGP framework encompasses both State obligations and corporate responsibilities, the focus of the Article is on the latter. The corporate responsibility to respect constitutes a non-binding "global standard of expected conduct" applicable to business enterprises, which exists independently of States' abilities and/or willingness to fulfil their own human rights obligations.[114] In particular, businesses are expected to "avoid causing or contributing to adverse human rights impacts through their own activities, and address such impacts when they occur," as well as "seek to prevent or mitigate adverse human rights impacts that are directly linked to their operations, products or services by their business relationships, even if they have not contributed to those impacts."[115] As such, the corporate responsibility to respect stems from both a social expectation, sometimes referred to as a company's "social licence to operate,"[116]

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

112. *Id*. at 4.

113. *Id*.

114. UNGP, annex para. 11.

115. UNGP, annex para. 13.

116. John Ruggie (Special Representative of the Secretary-General), *Protect, Respect and Remedy: A Framework for Business and Human Rights*, para. 54, U.N. Doc. A/HRC/8/5 (Apr. 7, 2008).

as well as the existence of a link between a company's activities and adverse human rights impacts.

While some human rights may be at greater risk and therefore require heightened attention in particular industries or contexts, the responsibility to respect applies to *all* human rights on the basis that "business enterprises can have an impact—directly or indirectly—on virtually the entire spectrum of these rights."[117] The responsibility to respect also applies to *all* business enterprises "regardless of their size, sector, operational context, ownership and structure."[118]

A frequently aired concern is that the corporate responsibility to respect may risk over-burdening profit-driven companies and chilling innovation.[119] Under the UNGP framework, however, the scale and complexity of the means through which businesses meet their responsibility to respect may vary according to the size, sector, operational context, ownership and structure, as well as the severity of an enterprise's adverse human rights impacts judged according to their "scale, scope and irremediable character."[120] Moreover, as Emily Laidlaw has explained, while operationalizing the corporate responsibility to respect inevitably entails a degree of market disruption in order to realign business conduct along human rights-compatible terms, the aim is "to narrowly tailor the obligations to minimize disruption beyond the intended purpose of encouraging human rights compliance."[121]

1. The Value of a Human Rights-Based Approach to Platform Moderation

Applied to the specific context of online platform moderation, the value of a human rights-based approach is threefold. First, a human rights-based approach provides online platforms with an organizing framework to transform their predominantly *ad hoc* and reactive approaches to the development of platform moderation policies towards a more *principled and structured approach*.[122] To satisfy their responsibility to respect

117. UNGP, annex para. 12.
118. *Id.* Principle 14 & Commentary.
119. LAIDLAW *supra* note 4, at 242 (critiquing this concern).
120. UNGP, Principle 14 & Commentary.
121. LAIDLAW, *supra* note 4, at 242.
122. *See* SUZOR, *supra* note 108, at 173.

human rights, online platforms are expected to put in place a range of policies, processes and procedures appropriate to their size and circumstances, which should include at a minimum: a high-level policy commitment to meet their responsibility to respect human rights; [123] a human rights due diligence process that identifies, prevents, mitigates and accounts for actual and potential human rights impacts of their activities;[124] verification of whether adverse human rights impacts are being addressed by tracking the effectiveness of company responses, whilst communicating relevant policies and processes externally to affected stakeholders; [125] and appropriate remediation of any adverse human rights impacts they cause or to which they contribute.[126] In this way, a human rights-based approach offers a structured methodology for approaching the development of platform moderation systems, as well as guidance concerning the types of measures through which the responsibility to respect can be operationalized in practice.[127]

Second, a human rights-based approach also provides platforms with the tools to assess the actual and potential human rights impacts of their platform moderation rules, processes, and procedures *holistically*, spanning their conception, design, and testing, their deployment in different contexts, and their ongoing monitoring and evaluation.[128]

Finally, a human rights-based approach provides platforms with *a common conceptual language* to identify the impact of their moderation rules, processes and procedures in different contexts and to explain, discuss, and justify their moderation decisions in an open and transparent manner.[129] To this end, international human rights law establishes thresholds for when rights have been interfered with, together with a series of tests to determine when

123. UNGP, Principle 16 & Commentary.

124. *Id*. Principles 17-19 & Commentary.

125. *Id*. Principles 20-21 & Commentary.

126. *Id*. Principles 22, 29 & 31 and Commentary UNGP.

127. *See* McGregor, Murray, & Ng, *supra* note 90, at 313 & 329-35; Amnesty International & Access Now, *The Toronto Declaration: Protecting the Right to Equality and Non-discrimination in Machine Learning Systems*, (May 16, 2018) [hereinafter Toronto Declaration, para. 42-56.

128. *See* McGregor, Murray, & Ng, *supra* note 90, at 325, 327-28, 334.

129. SUZOR, *supra* note 108, at 192.

rights may be restricted.[130] Importantly, international human rights law does not always dictate a specific or uniform outcome, but it provides a framework and vocabulary for platforms to assess whether the human rights impacts of their moderation systems are justifiable—with due sensitivity to the objectives and interests of the communities they nurture and the different contexts in which they operate.[131]

2. Limitations and Challenges to a Human Rights-Based Approach to Platform Moderation

Notwithstanding its value, it is important not to view a human rights-based approach to platform moderation as a panacea.[132] In particular, the approach is subject to at least three limitations and challenges. First, a human rights-based approach is *not a silver bullet for alleviating all harms* that arise on online platforms. Given the sheer volume of content moderated by platforms, a degree of human and algorithmic error is unavoidable.[133] Moreover, content moderation inevitably involves trade-offs between competing rights and interests. As such, the aim of a human rights-based approach is not to resolve moderation trade-offs in ways that are attractive to everyone, but more modestly to reduce adverse human rights impacts and more openly and transparently manage the trade-offs between the different rights and interests inevitably implicated by platform moderation practices.

Second, since the corporate responsibility to respect is non-binding, there is also *the challenge of enforcement*. While a human rights-based approach can be implemented on a purely self-regulatory basis, it is likely that platforms will be resistant where their commercial interests and profitability are threatened. With this in mind, a combination of social pressure and smart governmental (co-)regulation is likely to be required to assist and incentivize platforms to ensure the human rights compatibility of

130. McGregor, Murray, & Ng, *supra* note 90, at 326.

131. *See* SUZOR, *supra* note 108, at 198-201.

132. *See* McGregor, Murray, & Ng, *supra* note 90, at 313.

133. *See* M. Masnick, *Impossibility Theorem: Content Moderation at Scale is Impossible to do Well*, TECH DIRT (Nov. 20, 2019), https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml [https://perma.cc/9JKU-WHSQ].

their content moderation systems. Yet, given the inherent limits of social pressure, [134] as well as the risk of heavy-handed governmental regulation, [135] effective enforcement of a human rights-based approach is far from guaranteed.[136]

Arguably the biggest challenge, however, resides in *the translation of general human rights principles into particular rules, processes and procedures tailored to the platform moderation context*. [137] This task is complicated by the distinct and highly variable capacities and functions of platforms compared to States,[138] the diversity of products and services offered by today's largest online platforms,[139] the nascent state of development of the scope and content of businesses' human rights responsibilities,[140] and the difficulty of defining how platforms should exercise their pseudo-judicial role of weighing competing rights and interests

134. A promising proposal that is currently gaining traction is the creation of voluntary multistakeholder Social Media Councils. *See generally* ARTICLE 19, *The Social Media Councils: Consultation Paper* (June 2019). On informal regulation of online content, see also Gorwa, *supra* note 54.

135. On the features of freedom of expression and online platforms that make designing a legitimate system of speech governance particularly difficult, see generally Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation*, *Aegis Series Paper No. 1903* (2019), at 4-8; Damien Tambini, *Rights and Responsibilities of Internet Intermediaries in Europe: The Need for Policy Coordination*, CENTRE FOR INTERNATIONAL GOVERNANCE INNOVATION (Oct. 28, 2019), https://www.cigionline.org/articles/rights-and-responsibilities-internet-intermediaries-europe-need-policy-coordination [https://perma.cc/Y5R2-HVA8].

136. It is, however, possible to envisage smart models of government regulation in this context. *See, e.g.,* Kaye Content Moderation Report, *supra* note 10, at 19 ("smart regulation"); Bunting, *supra* note 8, at 176 ("procedural accountability"); French Interim Report, *supra* note 23, at 17 ("a regulatory policy based on a compliance approach to be applied and designed with pragmatism and agility"); Douek, *supra* note 135, at 8 ("a model of 'verified accountability'"); and Marsden, Meyer, & Brown, 'Platform Values and Democratic Elections: How Can The Law Regulate Digital Disinformation', *Computer Law & Security Review* (*forthcoming*).

137. *See* LAIDLAW, *supra* note 4, at 233.

138. SUZOR, *supra* note 108, at 247-48 ("There's no easy answer yet about what different societies expect from digital media platforms. We wouldn't want even the largest platforms to be bound by the same rules that regulate state power").

139. *See, e.g.,* Ingram, *Talking with Former Facebook Security Chief Alex Stamos*, GALLERY BY CJR (Oct. 2019), https://galley.cjr.org/public/conversations/-LsHiyaqX4DpgKDqf9Mj [https://perma.cc/2VFM-NRDR]. (in which Alex Stamos observes how Facebook "is actually something like a dozen different products strung together", with each product possessing "very different safety, security and trust models" and "very different levels of amplification and therefore potential for abuse").

140. McGregor, Murray, & Ng, *supra* note 90, at 313.

across a wide range of societal contexts.[141] The remainder of the Article takes up this final challenge, exploring some of the choices and hurdles that online platforms are likely to confront in attempting to implement their corporate responsibility to respect in the content moderation context.[142]

B. The Substance of Content Moderation

Turning first to the substance of content moderation, adherence to a human rights-based approach requires online platforms to align their substantive moderation rules with international human rights law. The practice of content moderation potentially implicates a range of human rights, including rights to equality, non-discrimination, privacy, and fair process. However, the right that is impacted to a particularly significant extent in this context—and which forms the focus of discussion in this section—is the right to freedom of expression. Pursuant to Article 19(2) of the ICCPR, the right to freedom of expression is defined in broad terms to include the freedom "to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice." [143] Importantly, any restriction of an individual's right to freedom of expression must satisfy the tripartite test of legality, legitimacy, and necessity set out in Article 19(3) of the ICCPR.

Applying a human rights-based approach in practice, these provisions should be at the center of platform moderation policies. As David Kaye has argued, online platforms "should incorporate directly into their terms of service and 'community standards' relevant principles of human rights law that ensure content-related actions will be guided by the same standards of legality, necessity and legitimacy that bind State regulation of expression." [144] Yet, while the tests of legality, legitimacy, and necessity are relatively simple to elaborate in the abstract, their

141. LAIDLAW, *supra* note 4, at 112 & 243-44.

142. *See also* Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 *Duke L. & Tech. Rev.* 26 (2019); Council of Europe, 'Recommendation CM/Rec (2018)2 of the Committee of Ministers to Member States on the Role and Responsibilities of Internet Intermediaries', 7 Mar. 2018, at Appendix, para. 2.

143. Article 19(2) ICCPR.

144. Kaye Content Moderation Report, *supra* note 10, para. 45.

translation from the State to the corporate context of platform moderation is likely to pose a number of challenges in practice.

1. Legality

Arguably the simplest condition to translate to the platform moderation context is the requirement that restrictions on the right to freedom of expression must be "provided by law."[145] In General Comment No. 34, the UN Human Rights Committee ("HRC") confirmed that restrictions on the right to freedom of expression should be "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly" and "made accessible to the public."[146] Additionally, "unfettered discretion" should not be conferred on those charged with executing restrictions of freedom of expression, while "sufficient guidance" should be provided to enable such persons to determine "what sorts of expression are properly restricted and what sorts are not."[147] Although platforms are not empowered to make formal laws, there are a range of actions that online platforms could adopt to align their content moderation rules with these standards.

In terms of *accessibility*, platforms could clearly alert users to the existence of their terms of service and community standards documents both upon registration and during general use of their platforms. Platforms could also ensure such documents are easy to find on their sites and available in different languages in line with their global reach. In addition, platforms could maintain an accessible public archive of former versions of the moderation rules and ensure users are notified of any updates to their policies as well as the reasons for such changes.[148]

With respect to *precision*, today's leading online platforms already elaborate on the different categories of content that will be subject to moderation, often developing distinct standards for general user-generated content and promoted content such as

145. Article 19(3) ICCPR.

146. UN Human Rights Committee, General Comment No. 34: Article 19: Freedom of opinion and expression, U.N. Doc. CCPR/C/GC/34, 12 Sept. 2011 [hereinafter General Comment No. 34], para. 25.

147. *Id*.

148. *A Rights-Respecting Model of Online Regulation by Platforms*, GLOBAL PARTNERS DIGITAL (May 2018), https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf [https://perma.cc/BY2Y-7GJ9].

advertising. As noted earlier in the Article, however, concerns have arisen that such policies often lack clarity and specificity. Responding to these concerns, platforms could improve the level of detail provided with respect to categories of problematic content by producing more detailed guidance notes to accompany their moderation rules, as well as elaborating real or hypothetical examples and case studies to illustrate how such rules are applied in practice. [149] Beyond precision concerning the substance of moderation rules, platforms could also provide clearer information concerning how content is flagged, the different forms of stakeholder engagement conducted as part of the process of formulating particular rules, the range of response measures that may be implemented when different types of content are found to be in violation of community standards (for example, filtering, blocking, removal, deprioritization, demonetization, and/or suspension or termination of accounts), as well as any review and grievance procedures that are available to users.

Twitter's recent update concerning when tweets will be allowed to remain on the platform despite violating the platform's rules because they are deemed to be in "the public interest" provides a clear example of how a platform can improve the precision of its moderation policies in practice. [150] The update includes details concerning who the policy applies to, what criteria will be used to define the public interest (including, for example, the immediacy and severity of potential harm from the rule violation), as well as what action will be taken in response to such tweets (including, for example, placing a notice on such tweets to provide additional context and clarity). The update brings welcome detail to a policy that has traditionally been shrouded in mystery. At the same time, since the criteria used to define the public interest will often require difficult judgments, Twitter could further align its policies with the standards required by the legality

149. Kaye Content Moderation Report, *supra* note 10, at 15; GLOBAL PARTNERS DIGITAL, *supra* note 148, at 16.

150. Twitter Safety, *Defining Public Interest on Twitter*, TWITTER BLOG (June 27, 2019), https://blog.twitter.com/en_us/topics/company/2019/publicinterest.html [https://perma.cc/NS9G-FVRF]. For an additional example concerning the alignment of platform policies on hate speech with the legality standard under international human rights law, *see generally* Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, UN Doc A/74/486, Oct. 9, 2019 [hereinafter Kaye Hate Speech Report], para. 4-28 & 46-50.

test by accompanying these criteria with case studies to further assist users to understand how the policy is applied in practice.

2. Legitimacy

Article 19(3) of the ICCPR also provides that any restriction of the right to freedom of expression must pursue one of a limited number of legitimate aims, namely: respect for the rights or reputations of others; the protection of national security; the protection of public order; the protection of public health; or the protection of public morals. According to the UN HRC, "restrictions must be applied only for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicated."[151]

There are also certain exceptional types of expression that States are required to prohibit under international law.[152] Article 20 of the ICCPR, for example, establishes an obligation to prohibit propaganda for war and any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence—a prohibition which, according to the UN Special Rapporteur on Freedom of Opinion and Expression, should be understood to apply to a broader set of protected categories now covered under international human rights law, including "race, colour, sex, language, religion, political or other opinion, national or social origin, property, birth or other status, including indigenous origin or identity, disability, migrant or refugee status, sexual orientation, gender identity or intersex status."[153] Beyond the ICCPR, Article 4 of the Convention on the Elimination of Racial Discrimination establishes an obligation to prohibit all dissemination of ideas based on racial superiority or hatred and incitement to racial discrimination, while Article 3(1) of the

151. General Comment No. 34, *supra* note 146, para. 22.

152. *See generally* Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, UN Doc A/66/290, 10 Aug. 2011 [hereinafter La Rue Report], para. 20-36.

153. Kaye Hate Speech Report, *supra* note 150, para. 9. *See also* Van Ho, *Twitter's Responsibility to suspend Trump's, and Rouhani's, Accounts, Part 1*, OPINIO JURIS (Jan. 21, 2010), https://opiniojuris.org/2020/01/21/twitters-responsibility-to-suspend-trumps-and-rouhanis-accounts-part-1/ [https://perma.cc/HK5Z-V2TH]; Van Ho, *Twitter's Responsibility to Suspend Trump's, and Rouhani's, Accounts*, *Part 2*, OPINIO JURIS (Jan. 21, 2020), https://www.newsbreak.com/news/0Nu8Hwu2/twitters-responsibility-to-suspend-trumps-and-rouhanis-accounts-part-2 [https://perma.cc/UZ68-VDAN].

Optional Protocol to the Convention on the Rights of the Child on the sale of children, child prostitution and child pornography establishes an obligation to prohibit the production, distribution and dissemination of child pornography. In addition, Security Council resolution 1624 (2005) calls upon States to "prohibit by law incitement to commit a terrorist act or acts," while certain forms of expression are also prohibited under international criminal law including, for example, direct and public incitement to commit genocide.[154]

Applied to the platform moderation context, two implications flow from these provisions.[155] First, platforms should prohibit the same exceptional and narrowly-defined forms of expression that States are required to prohibit under international law. Importantly, since the prohibition of these categories of content amount to restrictions on the right to freedom of expression, their prohibition must comply with the tripartite test of legality, legitimacy, and necessity set out in Article 19(3) of the ICCPR.[156] A particular challenge in this regard resides in defining these forms of expression, several of which – such as incitement to commit terrorism – have been subject to conflicting guidance and interpretation.[157] The importance of this challenge should not be understated: if platforms define these categories too broadly, they risk removing content beyond what is necessary in pursuance of a legitimate aim; by contrast, if platforms define these categories too narrowly, they risk nurturing dangerous online environments. I return to this challenge below in the context of examining the test of necessity.

Second, a platform should only restrict other categories of content provided such restrictions are in pursuance of at least one

154. *See generally* GREGORY S. GORDON, ATROCITY SPEECH LAW: FOUNDATION, FRAGMENTATION, FRUITION (2017).

155. *See* GLOBAL PARTNERS DIGITAL, *supra* note 148, at 16-17

156. La Rue Report, *supra* note 152, para. 37; General Comment No. 34, *supra* note 146, para. 50-52. See, however, OSCE Representative on Freedom of the Media, *Propaganda and Freedom of the Media*, Non-Paper (2015), at 15-17 (arguing that freedom of expression under the ICCPR "should be interpreted as not including war propaganda and hate speech that constitutes incitement to discrimination, hostility or violence"); Van Ho, Part 2, *supra* note 153 ("The ICCPR's Article 20 prohibition stands in contrast to the ICCPR's Article 19(3) balancing terms for freedom of expression generally because the international community has determined that propaganda for war serves no public interest and cannot be favourably balanced").

157. La Rue Report, *supra* note 152, para. 20-36.

of the legitimate aims elaborated in Article 19(3). The prospect of applying this test in the platform moderation context has led to questions concerning the appropriateness of requiring platforms to justify their online content restrictions solely on the limited public interest grounds recognized under Article 19(3). [158] Although understandable, such concerns should not be overstated.

It is true that companies are not well-placed to assess threats to "national security" or "public order"—grounds that should be relied upon by platforms only on the basis of legal orders from States, which themselves are subject to the tripartite test set out in Article 19(3). [159] Nonetheless, many of the most common categories of content restricted by online platforms correspond with little difficulty to at least one of the legitimate aims elaborated in Article 19(3). [160] As David Kaye has observed, human rights standards "would justify taking action against anti-vaccination sites that harm public health, white supremacists who incite harm of others, and terrorist groups like ISIS that use platforms to extend their violence."[161]

In addition, it is suggested that the legitimate aim of "the rights of others" can afford platforms sufficient leeway to tailor their moderation policies to the different communities they are designed to serve.[162] In particular, reliance on the legitimate aim of "the rights of others" would enable a platform to examine whether restrictions on freedom of expression are necessary to create positive and supportive spaces that nurture the freedom of expression of specific categories of users (for example, children or those with mental health issues),[163] to protect the rights of users to privacy and security,[164] or even simply to create particular online experiences (for example, designing a platform for sharing dog

158. Aswad, *supra* note 142, at 52-56.

159. Kaye Hate Speech Report, *supra* note 150, para. 47(b).

160. GLOBAL PARTNERS DIGITAL, *supra* note 148, at 16-17.

161. David Kaye, *The Clash Over Regulating Online Speech*, SLATE (June 6, 2019), https://slate.com/technology/2019/06/social-media-companies-online-speech-america-europe-world.html [https://perma.cc/KG9F-QDJ3].

162. *See* General Comment No. 34, *supra* note 146, para. 28 ("the term "rights" includes human rights as recognized in the Covenant and more generally in international human rights law").

163. GLOBAL PARTNERS DIGITAL, *supra* note 148, at 17. *See also* KATE JONES, ONLINE DISINFORMATION AND POLITICAL DISCOURSE: APPLYING A HUMAN RIGHTS FRAMEWORK 29, 46 (2019).

164. Kaye, *supra* note 73, at 120.

photos to the exclusion of photos related to other pets) as part of the platform's entrepreneurial freedom to design, innovate, and conduct a business—the latter freedom falling within the platform's right to property and freedom of expression.[165]

At the same time, it is important to emphasize that the legitimate aim of "the rights of others" is not unlimited: it would not justify the establishment of platforms designed to serve communities whose purpose is to share content prohibited under international human rights law, such as revenge porn or racist sites. Moreover, the legitimate aim of "the rights of others" remains subject to the test of necessity, which, as the next section explains, generally requires a careful contextually-informed balancing of different rights and interests. In practice, therefore, the trickiest challenge for platforms seeking to align their substantive content moderation rules with human rights standards will not generally reside in identifying a relevant legitimate aim, but rather in weighing and balancing competing rights and interests in accordance with the final requirement under Article 19(3)—the test of necessity.

165. *See, e.g., Magyar Tartalomszolgáltatók Egyesülete & Index.hu ZRT v. Hungary*, HUDOC, at 10 (2016), http://hudoc.echr.coe.int/eng?i=001-160314 [https://perma.cc/E6EQ-YFYD] (referring to a company's right to freedom of expression); Rep. of the Special Rapporteur on the Promotion & Protection of the Right to Freedom of Opinion & Expression on Its Thirty-Second Session, U.N. Doc. A/HRC/32/38, para. 55 (2016) ("[i]t remains an open question how freedom of expression concerns raised by design and engineering choices should be reconciled with the freedom of private entities to design and customize their platforms as they choose."). *See also* WOLFGANG BENEDEK & MATTHIAS KETTEMANN, FREEDOM OF EXPRESSION AND THE INTERNET 106 (2013) (referring to "the right to property and the right to keep privately owned social networks private and subject only to terms of service"); Nicolas Suzor, *The Role of the Rule of Law in Virtual Communities*, 25 BERKELEY TECH. L. J. 1817, 1853-54 (2010) (referring to "the free speech interests of the providers"); Rikke Frank Jørgensen and Anja Møller Pedersen, *Online Service Providers as Human Rights Arbiters*, *in* THE RESPONSIBILITIES OF ONLINE SERVICE PROVIDERS 179, 183 (2017) (referring to "the freedom of the intermediary to conduct a business (provide internet services)"); Jack Balkin, *Virtual Liberty: Freedom to Design and Freedom to Play in Virtual Worlds*, 90 VA. L. REV. 2043, 2080 (2004) (referring to "the platform owner's constitutional right to design"); French Interim Report, *supra* note 23, at 19 (recognizing social networks' entrepreneurial freedom, including the right to define and apply terms of use, to exercise an unrestricted information ordering system and to innovate (especially for smaller operators).

3. Necessity

Article 19(3) of the ICCPR also provides that restrictions on the right to freedom of expression must be "necessary" for the achievement of at least one of the prescribed legitimate aims.[166] According to the UN HRC, the concept of necessity requires that restrictions on freedom of expression "not be overbroad," but rather represent "the least intrusive instrument amongst those which might achieve their protective function."[167] In addition, necessity requires "demonstrat[ing] in specific and individualized fashion the precise nature of the threat, and the necessity and proportionality of the specific action taken, in particular by establishing a direct and immediate connection between the expression and the threat."[168] In practice, the test of necessity generally requires a context-sensitive balancing of competing rights and interests.[169] However, applying the balancing assessments conducted in existing human rights caselaw to the platform moderation context is complicated for two reasons.

First, although a variety of sources may assist platforms in applying the test of necessity—including the concluding observations, commentaries and jurisprudence of the UN treaty bodies, the jurisprudence of international, regional and national courts, as well as reports produced by UN Special Rapporteurs and civil society groups—the guidance produced by these sources has not always been clear or consistent. For example, Article 20(2) of the ICCPR—which, it will be recalled, obliges States to prohibit any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence—has been subject to reservations issued by numerous States, divergent implementation in domestic law and practice, as well as

166. Article 19(3), ICCPR

167. General Comment No. 34, *supra* note 146, para. 34.

168. *Id.* para. 35.

169. *See, e.g.*, *Magyar Tartalomszolgáltatók Egyesülete and Index.hu ZRT v. Hungary*, HUDOC, para. 58 (2016), http://hudoc.echr.coe.int/eng?i=001-160314 [https://perma.cc/JDE9-3G3Y] (observing that the Court may be required to ascertain whether the domestic authorities have struck "a fair balance" when protecting two values guaranteed by the Convention which may come into conflict with each other in certain cases). *See also* Equality & Hum. Rts. Comm'n, *Guidance – Legal Framework: Freedom of Expression*, at 6 (2015), https://www.equalityhumanrights.com/sites/default/files/20150318_foe_legal_framework_guidance_revised_final.pdf [https://perma.cc/9359-JKP9].

inconsistent interpretation within international and regional jurisprudence.[170] Where such disagreement exists, platforms should be afforded greater leeway to determine the appropriate way to apply the relevant standards within their community standards, giving due consideration to the diversity of guidance available and the different contexts in which the platform operates, whilst also taking care so far as feasible to consult with local stakeholders.[171]

Second, even where guidance is clear and consistent, transplanting assessments of necessity from existing human rights case law to the platform moderation context is also complicated by the distinct and variable capacities and functions of online platforms compared to States, as well as the diversity of contexts in which platforms operate.[172] Moreover, since certain larger platforms, such as Facebook, offer a number of different services (for example, enabling users not only to post but also to promote and amplify their content through advertising) and manage a diversity of spaces (for example, the news feed, public pages, a

170. *See generally* Rabat Plan of Action on the Prohibition of Advoc. of Nat'l, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence, at Appendix, U.N. Doc. A/HRC/22/17/Add.4, (2012) [hereinafter Rabat Plan of Action]; ARTICLE 19, *Hate Speech Explained: A Toolkit,* at 70-75, (2015), https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf [https://perma.cc/27Q6-GHX9]; *see generally* Amal Clooney & Philippa Webb, *The Right to Insult in International Law*, 48 COLUM. HUM. RTS. L. REV. 1 (2017); Catherine O'Regan, *Hate Speech Online: An (Intractable) Contemporary Challenge?*, 71 CURRENT LEGAL PROBS. 403, 407 (2018).

171. *See also* LAIDLAW, *supra* note 4, at 241; Evelyn Aswad, *The Role of U.S. Technology Companies as Enforcers of Europe's New Internet Hate Speech Ban*, COLUM. HUM. RTS. L. REV. ONLINE 1, 13 (2016) http://hrlr.law.columbia.edu/hrlr-online/the-role-of-u-s-technology-companies-as-enforcers-of-europes-new-internet-hate-speech-ban/ [https://perma.cc/KS3V-YNCB] (arguing that, for the purpose of adhering to the UNGP, technology companies should adhere to international human rights standards under the ICCPR rather than regional human rights standards).

172. *See also*, Nicolas Suzor, *supra* note 165, at 1865 ("It is not possible to provide any definitive answers as to which values should be read into virtual community governance structures… The exact content and boundaries of any such limits will always be highly contextual"); Patrick Leerssen, *Cut Out by The Middle Man: The Free Speech Implications of Social Network Blocking and Banning In The EU,* 6 J. INTELL. PROP. INFO. TECH. & ELECTRONIC COM. L. 99, 112 (2015) (arguing that the balance between a ground for removing content and the end user's free speech rights "requires a different calculus than that which is applied to state interference…"); JONES, *supra* note 163, at 30 ("Establishing how existing norms apply in new contexts is likely to be contested, and reaching settled views takes time whether it is done through expert opinion, through the drafting of normative guidance, through state negotiation or litigation").

variety of groups, and personal profiles), the assessment of whether a restriction is necessary may vary across the services and spaces operated within the same platform.[173]

Notwithstanding the complexity of the task, it is nonetheless possible to identify a number of general principles within existing human rights jurisprudence that may assist online platforms in applying the test of necessity in practice. The following is not intended to be exhaustive. Rather, the aim is to illustrate the different types of considerations that platforms should consider in applying the test of necessary in the platform moderation context.

a. Local Context

First, assessments of necessity require careful consideration of *various dimensions of the local context where the content is transmitted*, including both the timing and location of the expression. In *Kim Jong-Cheol v. Republic of Korea*, for example, the UN HRC confirmed that it might be legitimate for a State to restrict the publication of political polling for a limited period in advance of an election in order to maintain the integrity of the electoral process.[174] In reaching this conclusion, the UN HRC expressly took into account "the recent historical specificities of the democratic political processes" in the Republic of Korea, including the vulnerability of its election culture and climate to political manipulation and irregularities.[175] In a similar vein, David Kaye has emphasized the importance of contextual considerations for determining whether an expression constitutes incitement to hatred, including factors such as "the existence of patterns of tension between religious or racial communities, discrimination against the targeted group, the tone and content of the speech, the person inciting hatred, and the means of disseminating the expression of hate."[176]

Applied to platform moderation, these findings suggest that platforms should consider whether certain restrictions may only

173. *See* Ingram, *supra* note 139.

174. Kim Jong-Cheol v. Republic of Korea, Hum. Rts. Committee, Comm. No. 968/2001, U.N. Doc. CCPR/C84/D/968/2001, para. 8.3 (2005).

175. *Id*.

176. Rep. of the Special Rapporteur on the Promotion & Protection of the Right to Freedom of Opinion & Expression on Its Sixty-Seventh Session, UN Doc A/67/357, para. 46 (2012). *See also* Rabat Plan of Action, *supra* note 170, para. 29.

be necessary during certain periods of time and how restrictions should apply in particular contexts, having due regard for social, historical, cultural and linguistic nuance as far as possible.[177] In the latter regard, it is notable that Samidh Chakrabarti, a director of product management at Facebook, recently confirmed that a key challenge for the company is "how not to think of our platforms as one thing that's the same across the world, but how should they be different in different regions to try to mitigate […] risks."[178] To meet this challenge, platforms should engage – so far as possible in light of their of size and circumstances— with local stakeholders to assist in the development of their content moderation rules, as well as any accompanying guidelines for their interpretation in the different contexts in which they operate.[179]

b. Platform Characteristics

Second, assessments of necessity also require consideration of *the means used to transmit the expression*, taking into account both the purpose of the expression and its intended audience. In the case of *Jersild v. Denmark*, for example, the European Court of Human Rights concluded that Denmark had violated a journalist's right to freedom of expression by convicting him for aiding and abetting the dissemination of racist remarks through the broadcast of a programme that included an item on young extremists.[180] In reaching this conclusion, the Court had regard not only to the manner in which the feature had been prepared and its contents but also the purpose of the program and the setting in which it was broadcast, including the fact that "the item was broadcast as part

177. *See* AccessNow, *supra* note 109, at 7 ("Companies should not apply content moderation rules in a "one size fits all" fashion… [but] should take social, cultural and linguistic nuance into account, as much as possible."); ARTICLE 19, *supra* note 134, at 13 ("A degree of variation is inherent to international human rights law . . . fundamental rights are general principles, they are standards, and when they are translated into actual detailed rules through (judicial) approaches, there is unavoidably a certain margin of manoeuvre that comes into play.").

178. David Ingram, *Facebook's New Rapid Response Team has a Crucial Task: Avoid Fueling Another Genocide*, NBC NEWS, (June 20, 2019), https://www.nbcnews.com/tech/tech-news/facebook-s-new-rapid-response-team-has-crucial-task-avoid-n1019821 [https://perma.cc/9UUU-ESAR]. *Cf.* Keller, *supra* note 56, at 8 (discussing "platforms' operational preference for a single set of rules").

179. AccessNow, *supra* note 109, at 7.

180. *Jersild v. Denmark*, HUDOC, para. 37 (1994), http://hudoc.echr.coe.int/eng?i=001-57891[https://perma.cc/N8PF-59RX].

of a serious Danish news programme and was intended for a well-informed audience."[181]

Translated to the platform moderation context, these findings suggest that the test of necessity should be applied with due sensitivity to the function of a platform and the size and nature of the community it serves.[182] For example, a human rights-based approach would afford leeway to platforms to implement more restrictive moderation rules where necessary to create virtual spaces specifically designed to nurture and protect particular communities (for example, restricting content that could trigger anxiety or panic amongst users of a platform designed for those with mental health problems) or to establish particular online experiences (for example, restricting photos of cats on a platform designed solely for sharing and discussing photos of dogs) provided doing so does not generate disproportionate adverse human rights impacts within and/or beyond the platform.[183]

By contrast, where an online platform functions as a more general space for the free exchange of ideas—particularly one that has become a dominant and essential channel for public communication due to a dearth of viable alternative platforms that command similar network effects—the application of the necessity test will generally require a more nuanced approach.[184] For these

181. *Id*. para. 30-37.

182. *See* Balkin, *supra* note 165, at 2080 ("Everything depends on the nature of the virtual space that the platform owner has created"); Suzor, *supra* note 165, at 1852 (referring to the importance of "a thorough examination of the circumstances and social structure of the particular community").

183. *See also* GLOBAL PARTNERS DIGITAL, *supra* note 148, at 17 ("there may be situations where platforms have been (or may be) developed for a specific purpose, or for a particular community, which needs restrictions on certain content to ensure that the platform can meet the legitimate needs of its users"); York & McSherry, *supra* note 109 ("smaller platforms dedicated to serving specific communities may want to take a more aggressive approach. That's fine, as long as Internet users have a range of meaningful options with which to engage").

184. The relevance of a platform's dominance finds support in the jurisprudence of the European Court of Human Rights, which has circumscribed the obligations of States to respect and ensure the right to freedom of expression in a range of contexts based on whether applicants whose expression has been restricted by private actors have at their disposal viable alternative platforms to exercise their expression. *See, e.g.,* Animal Defenders International v. the UK, 57 Eur. Ct. H.R. Ap. No.48876/08 (2013) 21, 41-43 (concluding that the UK's prohibition of political advertising on television and radio did not violate Article 10 ECHR, based in part on the fact that "a range of alternative media were available to the applicant", including "radio or television discussion programmes of a political nature", as well as "non-broadcasting media including the print media, the

virtual spaces, a useful way of approaching the application of the necessity test is to distinguish between a platform's *gatekeeping* function, through which it determines the permissibility of content, and its *recommendation* function, through which it determines the visibility of content.[185]

With respect to the gatekeeping function, it is suggested that the necessity test will generally require larger market-dominant platforms to adopt a more inclusive and permissive approach to the moderation of content, pursuant to which the removal of only a limited range of narrowly defined categories of content is likely to be deemed necessary in practice. Similar to States, platforms managing larger more general online spaces are likely to find it difficult to justify the necessity of removing most forms of hate speech that are merely offensive, disturbing or shocking, but which do not rise to the level of threats of violence, harassment or assault against individually identifiable victims or the advocacy of discriminatory hatred constituting incitement to hostility, discrimination or violence.[186] At the same time, since larger virtual spaces also enable users to disseminate their content to wider audiences, a human rights-based approach also expects these

internet (including social media) as well as [. . .] demonstrations, posters and flyers"); Cengiz and Others v. Turkey, App. Nos. 48226/10 and 14027/11, Eur. Ct. H.R. (2015) para. 51-55 (observing that YouTube contained "specific information of interest to the applicants that is not easily accessible by other means" to reach the conclusion that no viable alternatives were available to the applicants to exercise their freedom to receive and impart information and ideas). *See also* Christina Angelopoulos et al., *Study of Fundamental Rights Limitations for Online Enforcement Through Self-Regulation,* Inst. for Info. L. 50 (2015) (noting that the "degree of dominance" of a company is relevant to a human rights analysis because, *inter alia*, the nature of the service can make it more difficult to abandon the service, for example "when the alternative services are very limited or are not of practical worth"); Leerssen, *supra* note 172, at 112 ("An important factor in determining the intermediary's discretion should be their degree of dominance as reflected in the ECHR's case law"); Keller, *supra* note 56, at 18 ("Many critics argue [. . .] that the platform ecosystem has created new forms of scarcity. Even if users can still speak on less-popular platforms, they argue, those may be inadequate because not enough other people are there to listen or respond").

185. Timothy B. Lee, *Alex Jones is a crackpot – but banning him from Facebook might be a bad idea*, ARS TECHNICA (August 6, 2018), https://arstechnica.com/tech-policy/2018/08/op-ed-alex-jones-is-a-crackpot-but-banning-him-from-facebook-might-be-a-bad-idea [https://perma.cc/Q48R-VSYG] (referring to platforms as "two separate products: a hosting product and a recommendation product"). *See also* Keller, supra note 56, at 25-26 (referring to the "rank but don't remove" model of platform moderation, "requiring major platforms to offer an uncurated, unranked service but preserving their discretion over the curated version").

186. *See* ARTICLE 19, *supra* note 170, at 18-23.

larger platforms to exercise particular vigilance and engage in structured and sustained engagement with local stakeholders to understand the coded language that may be relied upon in particular contexts,[187] both in identifying prohibited forms of hate speech constituting incitement to discrimination, hostility or violence and in ensuring its timely removal in order to protect individuals and communities that may be adversely impacted by such speech whether online and/or offline—particularly in environments experiencing heightened tension or conflict.[188]

With respect to a platform's recommendation function, by contrast, it is suggested that the necessity test will generally afford larger platforms much greater leeway to adopt a more hands-on approach to moderation, permitting reliance on a diversity of measures to reduce the visibility of a broader range of categories of content—for example, by adopting higher standards for content that is amplified or microtargeted as a paid advertisement or a sponsored post,[189] attaching a warning label to photos and videos that depict various forms of especially graphic or violent content, or down-ranking content that has been identified to be disinformation by an independent fact-checking organization. In other words, it is envisaged that larger platforms will be able to rely on their recommendation function to ensure their sites remain functional and attractive spaces to interact, [190] as well as to

187. *See also* Kaye Hate Speech Report, *supra* note 150, para. 50 ("Human evaluation [. . .] must be based on real learning from the communities in which hate speech may be found, that is, people who can understand the "code" that language sometimes deploys to hide incitement to violence, evaluate the speaker's intent, consider the nature of the speaker and audience and evaluate the environment in which hate speech can lead to violent acts").

188. On the dangers of platforms failing to remove unlawful categories of hate speech, see, e.g., EQUALITY LABS, FACEBOOK INDIA: TOWARDS THE TIPPING POINT OF VIOLENCE: CASTE AND RELIGIOUS HATE SPEECH (2019).

189. *See* Ingram, *supra* note 139 (in which Alex Stamos argues that, "[i]f you are allowing for amplification because of money or are financially supporting speech, there is a much larger [platform] responsibility"). *See also* Henry Farrell, *A Conservative YouTube Star Just Lost his Income Stream for Homophobic Slurs*, WASH. POST, (June 6, 2019), https://www.washingtonpost.com/politics/2019/06/06/conservative-youtube-star-just-lost-his-income-stream-homophobic-slurs-heres-what-happened-why [https://perma.cc/Y8M3-B47A].

190. *See* Kyle Langvardt, *Regulating Online Content Moderation*, GEO. L. J. 1353, 1363 (2018) ("Any attempt to protect online speakers from oppressive content moderation must simultaneously accommodate the content moderation that makes the Internet's 'vast democratic forums' usable – a delicate and difficult balance").

differentiate their respective virtual spaces in ways that reflect their commercial objectives, culture and feel.[191]

c. Least Intrusive Restrictive Measure

Third—and very much related to the latter discussion—the test of necessity also requires consideration of whether the imposition of a restrictive measure is *the least intrusive amongst those which might achieve their protective function*. In the case of *Ballantyne, Davidson and McIntryre v. Canada*, for example, the UN HRC concluded that a prohibition on commercial advertising in English with the aim of protecting the vulnerable position of the francophone minority within Canada "may be achieved in other ways that do not preclude the freedom of expression, in a language of their choice, of those engaged in such fields as trade."[192] For instance, the law could have required that advertising appears in both French and English.[193]

In the platform moderation context, when content that violates a platform's moderation rules has been identified, the terms of service and/or community standards generally identify a range of actions and sanctions that may be taken in response. In order to align their content moderation policies with the least intrusive standard, platforms could commit, so far as feasible in light of their size and circumstances, to diversify the range of restrictive measures that may be adopted in response to different types of content.[194] With respect to disinformation, for example, while its removal may be necessary for certain circumstances, such as where it is used to incite violence,[195] in general, it is possible to

191. SUZOR, *supra* note 108, at 198-99; ARTICLE 19, *supra* note 134, at 13. On the challenge of promoting and protecting diversity with respect to the recommendation function of online platforms, see generally Natali Helberger, Paddy Leerssen, & Max Van Drunen, *Germany Proposes Europe's First Diversity Rules for Social Media Platforms*, *LSE Media Policy Project*, (May 29, 2019), https://blogs.lse.ac.uk/medialse/2019/05/29/germany-proposes-europes-first-diversity-rules-for-social-media-platforms [https://perma.cc/X9GP-LG3V].

192. Ballantyne, Davidson, McIntyre v. Canada, Communications 359/1989 and 385/1989, Human Rights Committee, para. 11.4 (May 5, 1993).

193. *Id*.

194. *See generally* Land & Hamilton, *supra* note 79; Kaye Hate Speech Report, *supra* note 150, para. 51-52.

195. *See, e.g*., Lu, *Update on Myanmar*, FACEBOOK NEWSROOM (August 15, 2018) https://about.fb.com/news/2018/08/update-on-myanmar [https://perma.cc/J638-TML2] (observing how "in Myanmar, false news can be used to incite violence, especially

envisage a broader spectrum of response options focused on combatting its spread and influence, including educational initiatives focused on media literacy, counter-narrative and fact-checking collaborations, demonetization, as well as reductions in its visibility.[196]

Beyond diversifying response options, larger platforms, in particular, could also commit to enhancing the ability of users to control the types of content they view on their platforms. Enhanced user controls would serve two functions: first, enabling users to protect themselves from the abusive behavior of other users online; and second, enabling platforms to manage more permissive speech environments in the knowledge that their users are equipped with the means to select the personalized online experience they desire.

Some platforms have already begun empowering users to set their own filters and rules for what they see online. Twitter, for example, enables users to mute or block the accounts of other users,[197] while Facebook enables users to unfollow other users, pages and groups in their news feed so as to avoid seeing their content whilst remaining connected to the entities producing it.[198] More recently, Facebook has confirmed plans "to give people more control of what they see," initially by enabling users to decide whether to view less content that is close to the line of violating the platform's standards and in the future by providing users with flexible controls over categories like nudity where cultural norms and personal preferences vary considerably around the world.[199]

While these initiatives are welcome, several commentators have argued that online platforms could go much further in terms of the level of control they provide to their users. Timothy Garton Ash and his colleagues, for example, have proposed that Facebook should establish a range of new controls, including a "news feed

---

when coupled with ethnic and tensions" and that therefore Facebook would remove "misinformation that has the potential to contribute to imminent violence or physical harm").

196. For an overview of platform policies concerning disinformation, see generally ARTICLE 19, *supra* note 32, at 27-29.

197. *How to Control Your Twitter Experience*, TWITTER, https://help.twitter.com/en/safety-and-security/control-your-twitter-experience [https://perma.cc/Q2WU-RVWU].

198. Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 14-15.

199. Zuckerberg, *supra* note 103.

analytics feature" that would deconstruct the extent to which different types of content appear on users' news feeds, an "adopt a different point of view" function that would expose users to feeds with entirely different content to their own, and buttons or sliders that would enable users to control "whether they wish to see more content that cross-cuts against their political ideology, whether they wish to see more news, and whether they wish their News Feed to be curated at all, or if it should proceed chronologically."[200] In the latter regard, buttons or sliders might also be developed to enable users to determine their tolerance across a range of content categories including, for example, nudity or graphic violence.[201]

An even more ambitious proposal would involve companies opening up their platforms to allow third parties to develop "collective lenses" or "feed recipes"—essentially, customized user interfaces with their own bespoke content visibility and permissibility policies to which users could subscribe.[202] As Mike Masnick explains, this approach would "push the power and decision making out to the ends of the network, rather than keeping it centralized among a small group of very powerful companies" so as to enable users to choose the online experience or filtering setup they desire.[203] In addition, as Tarleton Gillespie argues, such an approach would also empower groups of users and independent organizations to collaborate "to help curate the platform landscape, in areas and around topics they are most invested and expert in, at a granularity and precision greater than the platform could by itself."[204] Interestingly, in December 2019, Jack Dorsey announced that Twitter would provide funding for a small independent team, called BlueSky, to develop an open and

200. Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 16-17. *See also The Invisible Curation of Content: Facebook's News Feed and our Information Diets,* WORLD WIDE WEB FOUNDATION (April 2018), https://webfoundation.org/research/the-invisible-curation-of-content-facebooks-news-feed-and-our-information-diets/ [https://perma.cc/TBY3-YDYA].

201. Keller, *supra* note 49, at 24.

202. GILLESPIE, *supra* note 3, at 199 ("collective lenses"); *see also* Columbia Journalism School, *Peter Zenger Lecture with Jonathan Zittrain*, YOUTUBE (Nov. 13, 2018) ("feed recipes"); Masnick, *Protocols Not Platforms: A Technological Approach to Free Speech*, KNIGHT FIRST AMENDMENT INSTITUTE (Aug. 21, 2019), https://knightcolumbia.org/content/protocols-not-platforms-a-technological-approach-to-free-speech [https://perma.cc/H8PP-B8G3] ("a protocols-based system").

203. *See generally* Masnick, *supra* note 133.

204. *See generally* GILLESPIE, IMPROVING MODERATION (2018).

decentralized protocol standard for social media—opening up a potential pathway for this type of initiative to be operationalized in practice.[205]

Yet, while enhancing user controls would result in a less top-down approach to platform moderation, it is also important to recognize the limits and challenges that administering such controls would entail.[206] First, any system of user controls will only be as effective as the technology on which it relies. Given the current limits of artificial intelligence, it is likely to be a long time before platforms are able to establish user controls that enable users to accurately control their viewing experience with respect to complex context-dependent categories of content.[207] Second, platforms will presumably still need to ensure that certain categories of content are beyond user control, for example, due to their illegality.[208] With this in mind, difficulties may arise in selecting which categories of content should be excluded from user control, defining the boundaries of excluded categories of content, and policing those lines in practice.[209] Finally, it also seems fair to assume that many users will simply lack the time, energy, or inclination to take advantage of controls delegated to them, with the result that the most significant question may become what default settings apply and, more specifically, whether default settings should vary according to geographical region.[210]

---

205. Mike Masnick, *Twitter Makes A Bet On Protocols Over Platforms*, TECHDIRT (Dec. 11, 2019), https://www.techdirt.com/articles/20191210/21054943552/twitter-makes-bet-protocols-over-platforms.shtml [https://perma.cc/S3BA-LSUD]. For a skeptical perspective, see Michael Kwet, *Can Twitter Ever Be Decentralized*?, SLATE (December 2019), https://slate.com/technology/2019/12/jack-dorsey-open-decentralized-twitter.html [https://perma.cc/U4YM-KX77].

206. *See* Langvardt, *supra* note 190, at 1380-83 (on the relative benefits and drawbacks of enhanced user controls on platforms); *see also* Masnick, *supra* note 202.

207. *See* Zuckerberg, *supra* note 103 ("We won't be able to consider allowing more content until our artificial intelligence is accurate enough to remove it for everyone else who doesn't want to see it").

208. *Id*. ("Of course, we're not going to offer controls to allow any content that could cause real world harm").

209. Langvardt, *supra* note 190, at 1381-82. See, however, Masnick, *supra* note 202 ("the reality is that these kinds of communities [around things like child exploitation content or other criminal activities, for example] are already forming – often on the dark web – and the way they are dealt with today is mostly via law enforcement […] There is little reason to think that in a protocol-focused world, this problem would be all that different than what currently exists").

210. Mark Zuckerberg, *Building Global Community*, FACEBOOK (Feb. 16, 2017), https://www.facebook.com/notes/mark-zuckerberg/building-global-

d. Protective Function of Restrictive Measure

Finally, the test of necessity also entails assessing *whether the restrictive measures taken in response to particular forms of expression actually fulfill their protective function*. Applied to the content moderation context, platforms could commit to assessing whether a particular response measure has any unintended negative consequences that may outweigh its protective benefits through an evidence-based approach.[211] According to Daphne Keller, for example, some of the policies adopted by online platforms to counter violent extremism "may cultivate precisely the attitudes and animosities that counter-radicalization efforts are supposed to prevent."[212] As Keller explains, "if suppressing propaganda from real terrorists comes at the cost of high over-removal rates for innocent Arabic-language posts or speech about Islam generally, the trade-off may be not only disrespectful and unfair but dangerous."[213] Where a response measure is found to restrict the freedom of expression of users without effectively furthering a legitimate purpose, platforms should commit to revising their policies accordingly.

Ultimately, the application of the test of necessity is one of the most challenging aspects of operationalizing a human rights-based approach to content moderation—and one which would benefit significantly from an independent multistakeholder mechanism to assist platforms in determining what restrictions are necessary in light of their diverse functions and contexts of operation. Yet, even in the trickiest cases, the test of necessity remains valuable to the extent that it requires platforms to openly explain and justify the trade-offs they inevitably have to make in a manner that surfaces the different rights and interests involved.

C. The Process of Content Moderation

In addition to aligning the substance of their content moderation rules with international human rights law, adherence

community/10154544292806634/ [https://perma.cc/32EY-9E6B] (suggesting that "the default will be whatever the majority of people in your region selected, like a referendum").

211. Aswad, *supra* note 142, at 51-52.

212. Keller, *supra* note 49, at 22.

213. *Id*. at 24.

to a human rights-based approach also requires online platforms to address the adverse human rights impacts of their moderation *processes*. Important processes in this context include the revision of terms of service and community standards, the management of systems of community and algorithmic flagging, the governance of human and algorithmic decision-making, the transparency of user-generated and advertising content, and the response of platforms to regulatory pressures.[214]

In order to identify, prevent, mitigate and account for how they address the adverse human rights impacts of their moderation processes, online platforms should establish a policy commitment to meet their responsibility to respect and carry out ongoing human rights due diligence.[215] The due diligence process should be initiated as early as possible in the development of new platform activities or relationships and be undertaken at regular intervals throughout the life of an activity or relationship.[216] As part of the process, platforms should put in place "policies and processes through which they can both *know* and *show* that they respect human rights in practice."[217]

In practice, there are three core stages to the process of human rights due diligence:[218] first, identifying and assessing any actual or potential adverse human rights impacts with which a platform may be involved either through their own activities or as a result of their business relationships;[219] second, taking effective action to prevent and mitigate adverse human rights impacts and tracking those responses to ensure they are being implemented optimally;[220] and finally, transparently communicating sufficient information externally about the platform's efforts to identify, prevent and mitigate adverse human rights impacts so that the adequacy of any response measures may be evaluated.[221] To elaborate on the challenges of applying these standards in the platform moderation context, this section examines four forms of transparency and oversight—rule-making, decision-making,

214. *See* SUZOR, *supra* note 108, at 201-20.
215. UNGP, Principles 16-17 and Commentary.
216. *Id.* Principles 17-18 and Commentary.
217. *Id.* Principle 21 and Commentary.
218. *See* Toronto Declaration, *supra* note 127, para. 44.
219. UNGP, Principles 17-18 and Commentary.
220. *Id.* Principles 19-20 and Commentary.
221. *Id.* Principle 21 and Commentary.

content and advertising, and regulatory compliance—that platforms should address as part of their human right due diligence processes in practice.

1. Rule-making

First, platforms should address their *rule-making* transparency and oversight.[222] Online platforms have often been criticized for failing to adequately consult with their users, civil society groups, or the general public concerning the development and revision of their moderation rules. According to Rebecca MacKinnon, for example, Facebook adheres to "a Hobbesian approach to governance in which people agree to relinquish a certain amount of freedom to a benevolent sovereign who in turn provides security and other services." [223] This approach is in tension with a human rights-based approach to content moderation. According to Principle 18 of the UNGP, human rights due diligence conducted by business enterprises should draw on "internal and/or independent external human rights expertise" and involve "meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation." [224]

In recent years, online platforms have taken steps to offer more opportunities for participation and feedback concerning their moderation rules. Facebook, for example, has stated that when its content policy team meet every few weeks to discuss potential changes to its moderation policies, they regularly invite outside experts and have begun to make the meeting minutes publicly available.[225] Facebook also launched a series of public events around the world entitled *Facebook Forums: Community Standards* to obtain public feedback on its policies directly.[226] In addition, Facebook has voluntarily submitted to three audits—a

222. Kaye, *supra* note 73, at 121 ("rulemaking transparency").

223. MACKINNON ET AL., *supra* note 47, at 408.

224. UNGP, Principle 18 and Commentary.

225. Zuckerberg, *supra* note 103.

226. Monika Bickert, *Publishing Our Internal Enforcement Guidelines and Expanding Our Appeals Process*, FACEBOOK (Apr. 24, 2018), https://about.fb.com/news/2018/04/comprehensive-community-standards/ [https://perma.cc/WU8Y-PRAF].

civil rights review of its internal operations, an investigation into whether the platform is biased against conservatives, and a human rights impact assessment of the company's presence in Myanmar—and established an independent election research commission to identify research topics and select independent researchers to examine them using Facebook data.[227]

While these developments have resulted in important recommendations and commitments concerning the platform's moderation rules and policies, [228] Facebook and other major platforms could go much further in ensuring more structured multistakeholder participation in the development and revision of their content moderation rules. In recent years, a number of proposals have been put forward by scholars and civil society groups that platforms could consider for this purpose. For example, platforms could adopt "notice-and-comment" procedures to obtain public feedback on proposed changes to their moderation policies,[229] appoint outside experts as "amici curiae" or in the form of an advisory panel to inform their policy decisions,[230] and/or support the creation of independent multistakeholder bodies to help ensure the compatibility of platform moderation policies with international human rights law. [231] While these proposals are varied—and would be subject to feasibility constraints according to the size and circumstances of a platform—they are premised on a shared belief that platforms would benefit from a more structured and sustainable approach to stakeholder engagement concerning their rule-making processes.

---

227. Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 18-19; *see also* Mark Zuckerberg, *Preparing for Elections*, FACEBOOK (Sep. 13, 2008). https://www.facebook.com/notes/mark-zuckerberg/preparing-for-elections/10156300047606634/ [https://perma.cc/3XAN-S4TF].

228. *See, e.g.*, *Facebook's Civil Rights Audit – Progress Report*, 9-13 (June 30, 2019), https://about.fb.com/wp-content/uploads/2019/06/civilrightaudit_final.pdf [https://perma.cc/4UPA-UNKA] (outlining recommendations to improve Facebook's white nationalism, hate speech and harassment policies).

229. Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 76 (2019).

230. *Id.*; Gillespie, *supra* note 18, at 214; Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 19-20.

231. *See, e.g.*, Gillespie, *supra* note 18, at 214 (Public Ombudsman); ARTICLE 19, *supra* note 134 (Social Media Councils); GLOBAL PARTNERS DIGITAL, *supra* note 148, at 26-28 (Independent Online Platform Standards Oversight Body).

2. Decision-making

Beyond rule-making, platforms should also address their systems of transparency and oversight concerning their human and algorithmic *decision-making* processes. In practice, this form of transparency and oversight has both quantitative and qualitative dimensions. *Quantitative* transparency and oversight refer to the statistical information disclosed by online platforms concerning their content moderation systems. Applying a human rights-based approach, platforms should disclose statistical information concerning the different categories of content removal requests they receive, the different types of actors that submit such requests, and the range of measures adopted in response. Moreover, to the extent feasible, platforms could also develop accuracy metrics, which enable platforms to calculate and publish error rates for human reviewers and algorithmic detection for different categories of violation,[232] as well as a range of metrics to evaluate the effectiveness of platform decision-making—for example, a metric that measures the virality of content found to violate moderation rules prior to its removal.[233]

Although online platforms have generally made incremental progress in the quality of their transparency reports, greater care and attention could be directed towards how content moderation statistics are disaggregated. In particular, platforms could distinguish between requests received through lawful channels such as court-orders, demands received from governments pursuant to a platform's terms of service, requests received from Internet referral units, requests received pursuant to voluntary arrangements such as the *EU Code of Conduct on Countering Illegal Hate Speech Online*, complaints submitted by private users, and proactive actions taken by platforms themselves including different forms of algorithmic decision-making.[234] To this end, platforms could draw guidance from the *Santa Clara Principles on Transparency and Accountability in Content Moderation*,[235] which set out the minimum level of detail that platforms should be

232. Report of the Facebook Data Transparency Advisory Group, April 2019 [hereinafter Facebook DTAG Report] at 16.

233. *Id*. at 18-28.

234. Kaye Content Moderation Report, *supra* note 10, at 16-17.

235. The Santa Clara Principles on Transparency and Accountability in Content Moderation (2018) [hereinafter Santa Clara Principles].

expected to disclose concerning their content moderation practices.[236]

Quantitative insights by themselves, however, are an inadequate means for assessing platform decision-making systems—for the simple reason that, without any form of independent verification, it is impossible to discern whether aggregate statistics accurately reflect the definitions and standards elaborated in a platform's moderation rules. As Evelyn Douek has put it, "[w]ithout verification or intelligibility, aggregate reports become a form of transparency theater, deployed to ward off calls for greater accountability."[237] With this in mind, applying a human rights-based approach, platforms should also establish *qualitative* forms of transparency and oversight, ideally through independent forms of verification and auditing that aim to identify and assess the actual and potential adverse human rights impacts of their systems of community and algorithmic flagging of potentially disallowed content, the different tiers of human and algorithmic review that such content is subjected to, and any algorithms relied upon to personalize user experience—making sure to integrate findings from their impact assessments through appropriate action, track the effectiveness of such measures, and communicate how they address adverse human rights impacts externally. In practice, appropriate action will depend on the adverse human rights impacts identified.

In terms of *community flagging*, for example, concerns have been raised about the practice of minority and vulnerable groups being targeted by coordinated mass reporting sprees of their accounts by groups that are politically or ideologically opposed to them. Appropriate action to guard against this practice could include treating reporting sprees as abusive behavior that is prohibited in platform moderation rules and placing limits on how many reports any single account can make in a day.[238]

With respect to *human review*, concerns have been raised about inadequacies in the number and cultural competency of moderators to accurately and effectively remove content that

236. *See generally* Facebook DTAG Report, *supra* note 232.

237. Douek, *supra* note 135, at 12.

238. Dia Kayyali, *Facebook's Name Policy Strikes Again, This Time at Native Americans*, EFF DEEPLINKS BLOG (Feb. 13, 2015), https://www.eff.org/deeplinks/2015/02/facebooks-name-policy-strikes-again-time-native-americans [https://perma.cc/3TZQ-FFVR].

violates platform moderation rules in different local contexts, including countries in the midst of turmoil or conflict, such as Myanmar, Sri Lanka, Libya, and the Philippines.[239] Appropriate action to address these concerns could include platforms committing to ensure adequate cultural and linguistic expertise across all markets in which they operate,[240] as well as establishing structured forms of engagement with local stakeholders, including early warning and emergency escalation functions, to enhance the ability of platforms to prevent their sites being instrumentalized for the promotion of violence.[241] Beyond the accuracy and effectiveness of human review, concerns have also been raised about the damaging work conditions and inadequate labor protections afforded to human reviewers, a particularly serious concern in light of the severe psychological toll that content moderation entails.[242] Appropriate action to address such concerns could include disclosing more information concerning the number, diversity, location, working conditions, support, and training put in place for human moderators, as well as committing to establish adequate labor protections in accordance with international human rights standards.[243]

Finally, in terms of *algorithmic decision-making*, a range of concerns have arisen over the opacity of algorithms, including the potential for algorithmic biases that adversely impact the human rights of different groups of users. Appropriate action in response to these concerns could take various forms. Recognizing the current limits of algorithmic decision-making, platforms could commit to ensuring that there will always be meaningful human involvement in their algorithmic decision-making processes and adequate safeguards in place in case an algorithm acts unpredictably.[244] Facebook, for example, has recently committed

239. Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 11.

240. Kaye Content Moderation Report, *supra* note 10, at 18.

241. *See, e.g.,* Ingram, *supra* note 178 (describing a new Facebook tool that allows approved civil society groups to flag problematic material in a way that is seen more quickly by the company than if a regular user had reported the material).

242. *See generally* ROBERTS, *supra* note 89.

243. Kaye Content Moderation Report, *supra* note 10, at 18.

244. GLOBAL PARTNERS DIGITAL, *supra* note 148, at 22; McGregor, Murray & Ng, *supra* note 90, at 341-42. *See also* Spandana Singh, *Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content*, NEW AMERICA (July 22, 2019), https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-

to removing thousands of targeting terms for advertisers offering housing, employment, or credit opportunities as part of a settlement of multiple discrimination lawsuits.[245] While a promising development, similar changes may be required for other categories of online advertising—for example, political advertising in light of evidence that voter suppression and intimidation tactics have been deployed on Facebook.[246]

Steps could also be taken to improve algorithmic transparency by explaining when algorithms are used, how they work, and their consequences for users in different contexts including the key criteria that underpin particular decisions.[247] In practice, however, enhancing algorithmic transparency in the platform moderation context raises a number of challenges. Meaningful transparency is complicated by the fact that algorithms are frequently altered—Google, for example, updates its algorithms hundreds of times per year—and often rely on machine learning techniques, which can lead to a divergence between what programmers believe an algorithm does and how it actually behaves.[248] There is also a risk that making algorithms more transparent will expose platforms to manipulation and gaming.[249] And even if a human moderator is "in the loop" of an algorithmic decision-making process, the risk remains that the moderator may unquestioningly or subconsciously defer to the algorithmic decision due to perceptions of technological neutrality and accuracy—often referred to as "automation bias."[250]

---

user-generated-content/ [https://perma.cc/7J2L-YNHF] (arguing that platforms should invest greater efforts in hiring developers who are non-Western and non-English speakers to help reduce data and creator biases).

245. *See generally* Facebook's Civil Rights Audit – Progress Report, (June 30, 2019), at 15-17. *Cf.* Ava Kofman & Ariana Tobin, *Facebook Ads Can Still Discriminate Against Women and Older Workers, Despite a Civil Rights Settlement*, PROPUBLICA (Dec. 13, 2019), https://www.propublica.org/article/facebook-ads-can-still-discriminate-against-women-and-older-workers-despite-a-civil-rights-settlement [https://perma.cc/QAJ5-GTPS].

246. Siva Vaidhyanathan, *Facebook is Ripe for Exploitation – Again – in 2020*, GUARDIAN (July 9, 2019), https://www.theguardian.com/commentisfree/2019/jul/09/facebook-is-ripe-for-exploitation-again-in-2020 [https://perma.cc/MF9Y-BXUW].

247. Kaye AI Report, *supra* note 84, at 18.

248. CoE Report, *supra* note 84, at 38.

249. Langvardt, *supra* note 190, at 1384.

250. McGregor, Murray, & Ng, *supra* note 90, at 338-41.

Some commentators have also cautioned that focusing narrowly on individual rights to algorithmic transparency risks creating a "transparency fallacy" since "individuals are mostly too time-poor, resource-poor, and lacking in the necessary expertise to meaningfully make use of these individual rights."[251] Edwards and Veale, for example, argue that "creating better systems, with less opacity, clearer audit trails, well and holistically trained designers, and input from concerned publics seems eminently more appealing than grimly pursuing against the odds a "meaningful" version of the interior of a black box."[252] Following this approach, platforms should devote more time and attention to establishing systems of internal and external review concerning how their algorithms are developed and deployed in decision-making processes.[253]

Proposals to expose algorithms to external audits are also likely to meet with resistance because platforms generally regard their underlying software code as protected proprietary technology.[254] However, these concerns are not insurmountable. Joshua Kroll, for example, has discussed the possibility of "accountability by design," which relies on techniques from computer science "to create systems with properties that can be checked by regulators or the public *without revealing the underlying code and data*."[255] Other options include examining the outcomes of algorithmic decision-making to identify biases or using counterfactual explanations to reveal how algorithms arrive at their decisions.[256] Ultimately, regardless of the precise modalities for improving the transparency and oversight of human and algorithmic processes, an important value of the human rights-based approach is the expectation it generates for platforms to adopt a cyclical approach to accountability whereby ongoing due

251. Lilian Edwards & Michael Veale, *Slave to the Algorithm? What a 'Right to an Explanation' is Probably Not the Remedy You Are Looking For*, 16 DUKE L. & TECH. REV. 18, 67 (2017).

252. *Id*. at 82.

253. Kaye AI Report, *supra* note 84, at 19-20; Douek, *supra* note 135, at 12.

254. CoE Report, *supra* note 84, at 38.

255. Joshua A. Kroll, *Accountable Algorithms (A Provocation)*, LSE MEDIA POLICY PROJECT (Feb.10, 2016), https://blogs.lse.ac.uk/medialse/2016/02/10/accountable-algorithms-a-provocation/ [https://perma.cc/XH86-7CD6].

256. *See generally* Sandra Wachter, Brent Mittelstadt, & Chris Russell, *Counterfactual Explanations without Opening the Black Box: Automated Decisions and the GDPR*, 31 HARV. J. OF L. & TECH. 841 (2018).

diligence efforts are enhanced by lessons learned from previous results across the entire human and algorithmic decision-making life cycle.[257]

3. Content and Advertising

The third form of transparency and oversight that platforms should address concerns the data disclosed to users about the *content and advertising* they view and share. Although content and advertising transparency on platforms has traditionally been very limited, improvements have been made in recent years. [258] Facebook, for example, enables users to access all advertisements that a page is running across Facebook, Instagram, and Messenger, to identify any recent name changes and the date a page was created, and to obtain some insight into the categories used by advertisers to micro-target users through their "Why am I seeing this?" button. In addition, in the US, Facebook requires all political and issue advertisements to make clear who paid for them, to be stored for up to 7 years in a public archive which anyone can access and search to identify how much was spent on an individual advertisement and the audience it reached, and also requires anyone running such ads to verify their identity and location.[259]

While these types of measures are a step in the right direction, they are still relatively modest. Tarleton Gillespie, for example, has challenged online platforms to provide more data about links that are shared by users, as well as more radical transparency concerning all types of advertising. With respect to links shared by users, Gillespie proposes a dashboard that would appear when hovering over a link, which would report "how long the source of that link has been online, a graph of how much and how quickly that headline is being forwarded, the headlines before and after this one from the same source, and how often other articles from that source have been disputed by fact checkers," as well as details

257. McGregor, Murray, & Ng, *supra* note 90, at 327-29.

258. *But see* Jeremy B. Merrill & Ariana Tobin, *Facebook Moves to Block Ad Transparency Tools – Including Ours*, PROPUBLICA, (Jan. 28, 2019), https://www.propublica.org/article/facebook-blocks-ad-transparency-tools [https://perma.cc/F62D-N77P].

259. Zuckerberg, *supra* note 227. For a critical discussion of platform ad archives, *see generally* Leerssen et al., *Platform Ad Archives: Promises and Pitfalls,*8 INTERNET POL'Y REV. (2019).

about the person who posted the link, including how long they have been on the platform, the last five articles they shared, whether they have previously been reprimanded by the platform and whether the user read the article before forwarding it.[260] Gillespie also argues that platforms could reveal significantly more data about advertisements, including the days on which an advertisement was delivered to any users, the targeting criteria, the number of users the advertisement was delivered to directly and reach indirectly, how much the advertiser paid for its circulation, and a link back to the advertiser's profile page, all of which could be stored in a public and searchable archive.[261] Since online platforms extract significant economic value from the data shared by and inferred from the behavior of their users, radical transparency is not only reasonable in this context but also a tool that may assist civil society groups, journalists, and regulators to monitor and hold platforms and advertisers to account.

4. Regulatory Compliance

A final form of transparency and oversight concerns how online platforms respond to both mandatory regulatory measures and informal regulatory pressures that may have adverse human rights impacts on their users. According to Principle 23 of the UNGP, while business enterprises should comply with all applicable laws wherever they operate, they should also "seek ways to honor the principles of internationally recognized human rights when faced with conflicting requirements."[262] In particular, if the domestic context renders it impossible for companies to fully satisfy their corporate responsibility, "business enterprises are expected to respect the principles of internationally recognized human rights to the greatest extent possible in the circumstances, and to be able to demonstrate their efforts in this regard."[263]

Examples of the types of measures that platforms may adopt in this context have been elaborated by the Global Network Initiative ("GNI"). The GNI is a multistakeholder alliance of companies, civil society groups and academic institutions whose

260. Gillespie, *supra* note 204, iv.
261. *Id*. vii-viii.
262. UNGP, Principle 23.
263. *Id.* Principle 23 and Commentary.

aim is to "protect and advance freedom of expression and privacy in the Information and Communications Technology (ICT) industry globally." [264] Since its inception, the GNI has been subject to significant criticism, with concerns raised over its inadequate corporate membership, insufficiently independent assessment process, and lack of a remedial mechanism.[265] Without diminishing the force of these concerns, the documents that underpin the GNI's work nonetheless provide useful guidance regarding how online platforms should respond when confronted by State demands to undermine the freedom of expression rights of their users.

According to the GNI's *Principles on Freedom of Expression and Privacy*, when confronted by national laws, regulations or policies that do not conform to international standards, "companies should avoid, minimize, or otherwise address the adverse impact of government demands, laws, or regulations, and seek ways to honor the principles of internationally recognized human rights to the greatest extent possible."[266] To this end, the GNI's *Implementation Guidelines* elaborate a number of tools that participating companies are required to rely upon in practice.[267] In particular, participating companies agree to encourage governments to be specific, transparent and consistent in demands, laws, and regulations that impact freedom of expression, as well as engage proactively with governments to reach a shared understanding of how government restrictions can be applied consistently with international human rights law. Participating companies also agree to require governments to follow established domestic legal processes when seeking to restrict freedom of expression, and to request clear written communications from the government that explain the legal basis for such restrictions. Finally, participating companies are also required to narrowly interpret government

264. The Global Network Initiative, Principles on Freedom of Expression and Privacy [hereinafter "GNI Principles"], at 1.

265. Jørgensen, *supra* note 6, at 262-63; LAIDLAW, *supra* note 4, at 104-10.

266. GNI Principles, *supra* note 264, at 2.

267. The Global Network Initiative, Implementation Guidelines for the Principles on Freedom of Expression and Privacy, at 8-10. *See also* Hilary Hurd, *How Facebook Can Use International Law in Content Moderation*, LAWFARE (Oct. 30, 2019), https://www.lawfareblog.com/how-facebook-can-use-international-law-content-moderation [https://perma.cc/GJL4-G6M6] (proposing that Facebook should require States to submit formal explanations of why and how their take-down requests comply with the tripartite test set out in Article 19(3) ICCPR).

restrictions and demands so as to minimize the negative effect on freedom of expression and, where appropriate, to challenge any restrictions or demands that appear overbroad within domestic courts.

Beyond adopting measures that seek to reduce the impact of overly-intrusive State requests and regulations, platforms should also address the extent to which they publicly disclose the different pressures exerted by States over their moderation practices. For instance, platforms could consult or hire ombudspersons to assess State content removal requests and identify requests that would result in the removal of content that is significant for public debate.[268] By providing meaningful transparency, online platforms may be able to help illuminate and generate public conversation about the extent to which particular forms of regulatory pressure may be undermining the freedom of expression rights of their users.[269]

D. The Procedure and Remediation of Content Moderation

However well-articulated and enforced a platform's moderation processes may be, mistakes will inevitably occur that generate adverse effects for the freedom of expression of platform users. Addressing this situation, Principle 22 of the UNGP confirms that companies "should provide for or cooperate in their remediation through legitimate processes."[270] More specifically, Principle 29 of the UNGP provides that business enterprises "should establish or participate in effective operational-level grievance mechanisms for individuals and communities who may be adversely impacted."[271]

Grievance mechanisms can take many forms, but are generally intended to perform two important functions:[272] first, to support the identification of adverse human rights impacts in business operations; and second, to enable grievances to be addressed and adverse impacts remediated early and directly by business enterprises. Importantly, Principle 31 elaborates a set of

268. Citron, *supra* note 54, at 1067-69.
269. *Id*.
270. UNGP, Principle 22.
271. *Id*. Principle 29.
272. *Id*. Principle 29 and Commentary.

effectiveness criteria that grievance mechanisms should reflect, namely that they should be legitimate, accessible, predictable, equitable, transparent, rights-compatible, a source of continuous learning, and based on engagement and dialogue.[273]

Applying these principles and criteria to the specific context of content moderation, there are three areas that online platforms should address to improve the effectiveness of their procedural and remedial processes.[274] The first area concerns *due process*. Adherence to a human rights-based approach to content moderation requires platforms to afford due process to affected users—including adequate notice and avenues for appeal. Yet, as Evelyn Douek has recently observed, "*due* process does not mean *perfect* process" and "what 'due process' means needs to be determined contextually."[275] The challenge, as Douek puts it, is to identify "what the 'due' in 'due process' means in the context of the scale of the online platforms."[276] A useful point of departure for this conversation is offered by the *Santa Clara Principles*.

In terms of notice, the *Santa Clara Principles* suggest that "companies should provide notice to each user whose content is taken down or account is suspended about the reason for the removal or suspension."[277] At a minimum, notices should contain, in a language understandable to the user, information sufficient to allow the identification of the removed content, the specific provision of a platform's moderation policies that has been violated, how the content was detected and removed, and an explanation of the process by which the user can appeal the decision as well as an indicative time frame and what remedies may be available if successful.[278] In terms of appeals, the *Santa Clara Principles* suggest that "[c]ompanies should provide a meaningful opportunity for timely appeal of any content removal or account suspension."[279] At a minimum, meaningful appeal means human review by a person or panel of persons not involved in the initial decision, an opportunity to present additional

273. *Id*. Principle 31.
274. SUZOR, *supra* note 108, at 220-25 (referring to "scalable due process").
275. Douek, *supra* note 135, at 8 & 10 (emphasis in original).
276. *Id*. at 9.
277. Santa Clara Principles, *supra* note 235.
278. *Id*.
279. *Id*.

information, and notification of the results of the appeal, including a statement of the reasoning to enable the user to understand the decision.[280] In addition, given the importance of context for assessing content, it would also be beneficial for reviewers to be granted access to more contextual information about the pieces of content that they review on appeal.[281] Moreover, particular care should be taken to ensure that any appeals process is clear and easy to use for the average user and not susceptible to being gamed or abused. The appeals mechanism administered by Amazon, for example, has recently been criticized for becoming "the ultimate weapon in the constant warfare of Marketplace," its rules and processes "so confounding that it's given rise to an entire industry of consultants."[282]

While the *Santa Clara Principles* offer a useful starting point, the challenge remains in defining what adequate due process means in the specific context of different types of platforms. According to Douek, for example, platforms should strive for a "systematic understanding of due process" that contextually calibrates the level of process that is afforded to affected users in practice, for instance "by more explicitly differentiating between the different categories of speech that content moderation implicates, accounting for the difficulty of making a correct decision in each category, the public importance of the underlying speech, and how people experience different kinds of decisions."[283] To this end, some form of oversight mechanism could be useful in providing platforms with "a forum for error explanation and more deliberate choices between trade-offs involved in any system design."[284]

To date, there are indications that oversight mechanisms may emerge at both the cross-platform and individual platform levels. At the cross-platform level, for example, civil society group Article 19 has proposed the establishment of a multistakeholder accountability mechanism for platform moderation in the form of

280. *Id*.

281. Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 13.

282. Josh Dzieza, *Prime and Punishment*, VERGE (Dec. 19, 2018), https://www.theverge.com/2018/12/19/18140799/amazon-marketplace-scams-seller-court-appeal-reinstatement [https://perma.cc/42VB-P574].

283. Douek, *supra* note 135, at 10.

284. *Id*.

multistakeholder Social Media Councils.[285] At the individual platform level, Facebook recently concluded a consultation concerning the establishment of a new Oversight Board for Content Decisions aimed at providing "a new way for people to appeal content decisions."[286] The creation of each of these bodies raises a host of issues centered on institutional design, including questions concerning whether such bodies will be global, regional or national, the breadth of their jurisdiction and powers, and the extent of their independence and authority. Nonetheless, each has the potential to serve as a useful mechanism to help platforms navigate the complex terrain of translating international human rights standards to the platform moderation context—whether through reviewing the compatibility of emblematic individual cases with international human rights law and/or providing general guidance on the compliance of platform processes and procedures with international human rights standards.[287]

In addition to due process, a second area for platforms to address is the question of *remedies*. In terms of remedies for wrongful removal of content or suspension/termination of user accounts, David Kaye has suggested that online platforms "should institute robust remediation programmes, which may range from reinstatement and acknowledgement to settlements related to reputational or other harms."[288] In a typical case, reinstatement of the content or the account may be the most effective remedy. Other possible remedies such as a public apology, guarantees of non-repetition, or compensation may also be appropriate depending on

285. *See generally* ARTICLE 19, *supra* note 134.

286. Facebook Newsroom, *Global Feedback and Input on the Facebook Oversight Board for Content Decisions* (June 27, 2019), https://about.fb.com/news/2019/06/global-feedback-on-oversight-board [https://perma.cc/99YJ-GZKG]. *See also* Evelyn Douek, *Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility*, 21 N.C. J. L. & TECH. 1, 2-3 (2019).

287. *See, e.g.*, ARTICLE 19, *supra* note 134, at 13 (acknowledging that any Social Media Council mechanism will need to afford platforms a "margin of appreciation" to allow the application of international human rights standards in specific national contexts and enable differentiation between different companies and their respective products); Facebook Newsroom, *supra* note 286 (noting that a general theme to emerge from Facebook's consultation was that the Oversight Board "will need a strong foundation for its decision-making, a set of higher-order principles – *informed by free expression and international human rights law* – that it can refer to when prioritizing values like safety and voice, privacy and equality") (emphasis added).

288. Kaye Content Moderation Report, *supra* note 10, at 18.

the circumstances.[289] In terms of sanctions for violating content moderation rules, Kaye also suggests that platforms should have "graduated responses according to the severity of the violation or the recidivism of the user."[290] Sanctions might include, for example, de-amplification, de-monetization, requiring suspended users to issue an apology in order to be reinstated, or compensation.[291] A particular challenge in this regard—and one which would benefit from further multistakeholder reflection—is the question of how to establish a process for providing an effective remedy to victims who suffer physical, psychological or reputational harm as a result of failures by platforms to remove particular types of content, without at the same time incentivizing online platforms to over-censor the content of their users.

Finally, platforms should also address *remediation transparency and stakeholder engagement*. In both the design and implementation of their remediation processes, online platforms should engage relevant stakeholders, including through quantitative transparency concerning the frequency, patterns and reasons for appeals, so that they are well-equipped to identify policies and processes in need of reform.[292] Accompanying quantitative transparency, David Kaye has also suggested that platforms should improve their qualitative "decisional transparency" by developing a public, accessible and easily searchable jurisprudence of "platform law."[293] And ultimately, platforms should commit to address and revise problematic aspects of their remediation processes identified by stakeholders in a timely manner.

IV. CONCLUSION

The governance of speech in the digital age depends to a significant degree on the policies and processes of online platforms. Ensuring that these platforms govern in the public interest has emerged as one of the most pressing challenges for freedom of expression in the twenty-first century. In light of the

289. GLOBAL PARTNERS DIGITAL, *supra* note 148, at 24.

290. Kaye Hate Speech Report, *supra* note 150, para. 54-55.

291. *Id*.

292. GLOBAL PARTNERS DIGITAL, *supra* note 148, at 24.

293. Kaye Content Moderation Report, *supra* note 10, at 19. *See also* Garton Ash, Gorwa, & Metaxa, *supra* note 2, at 12.

wide range of problems that have been identified with existing platform moderation practices, the Article has explored the extent to which a human rights-based approach may help to alleviate such concerns. In particular, the Article has demonstrated how the adoption of a human rights-based approach would mark a significant shift towards a more structured and principled approach to content moderation by providing platforms with a framework and the conceptual tools to holistically assess and address the adverse human rights impacts of their moderation rules, processes and procedures. At the same time, the Article has been careful not to present the human rights-based approach as a panacea—revealing the various challenges and choices that online platforms are likely to confront in operationalizing such an approach in practice.

The implementation of a human rights-based approach to content moderation is not simple, raising complex questions concerning how to translate general human rights standards into particular rules, processes, and procedures tailored to the platform moderation context. This task is complicated by the diversity of services and spaces that online platforms offer and the wide range of societies in which they operate. Given this complexity, the risk inevitably arises that online platforms may try to co-opt the vocabulary of human rights to legitimize minor reforms at the expense of undertaking more structural or systemic changes to their moderation processes. Moreover, even if a human rights-based approach were to be effectively implemented by online platforms, it is important to remember that such an approach is not a silver bullet for alleviating all concerns that have been raised concerning platform moderation practices. Given the complexity of content moderation, there will always be trade-offs that are disagreeable to some users, while a degree of human or algorithmic error is unavoidable.[294]

The power and influence of online platforms also suggest that they should not be considered the exclusive province of any single regulatory paradigm. Rather, multiple paradigms will be needed to address different dimensions of online platforms, including, for example, data protection law, electoral and advertising regulation,

294. *See also* GODWIN, *supra* note 73, at 175-78.

and antitrust and competition law.[295] Furthermore, while the Article has focused on the responsibilities of the companies that manage online platforms, it is important to emphasize that the protection of freedom of expression online necessarily entails addressing the responsibilities of the far broader range of actors that participate in the digital public sphere—including, for example, governments, political parties, data brokers, mass media organizations, and advertisers.

Finally, it should also be that the online information ecosystem is more diverse than the online platforms discussed in the Article, encompassing a broader range of technologies, the unique characteristics of which will require bespoke policy responses. For example, the architecture of messaging services such as WhatsApp—where activity consists of encrypted personal conversations and groups involving up to 256 people—makes it much harder to stem the flow of disinformation compared to the virtual spaces administered by online platforms.[296] Going forward, more attention will need to be directed towards the broader array of services and technologies relied upon for online communication around the world.

295. *See also* European Data Protection Supervisor, 'EPDS Opinion on Online Manipulation and Personal Data', Opinion 2/2018, 19 Mar. 2018 at 13*ff*.

296. Cristina Tardágulia, Fabrício Benevenuto, & Pablo Ortellado, *Fake News Is Poisoning Brazilian Politics. WhatsApp Can Stop It*, N.Y. TIMES (Oct. 17, 2018), https://www.nytimes.com/2018/10/17/opinion/brazil-election-fake-news-whatsapp.html [https://perma.cc/2XDC-WK8Z].