The Limits of International Law in Content Moderation

evelyn douek*

In remarkably short order, there has been growing convergence, especially in academia and civil society, around the idea that major social media platforms should use international human rights law (IHRL) as the basis for their content moderation rules. Even platforms themselves have begun to agree. But why have these legendarily growth-obsessed companies been so quick to voluntarily say they are jumping on this bandwagon? Afterall, advocates for incorporating IHRL into content moderation governance generally envision it operating as a constraint on social media platforms' operations. There are both encouraging and less encouraging explanations. For the glass half-full types, there is the straightforward explanation that perhaps these companies genuinely care about human rights. But there is also a less optimistic possibility: companies are embracing the terminology so readily because they know that, in reality, it will not act as much of a constraint at all. This is the prospect explored in this Article. This Article is a sympathetic critique of the contributions IHRL can make to content moderation, highlighting the very real limits of IHRL as a practical guide to what platforms should do in many, if not most, difficult cases. It surveys the many arguments in favor of IHRL as a basis for content moderation rules. Ultimately, however, it argues that failing to acknowledge the considerable limitations of IHRL in this context will only serve the interests of platforms rather than their users by giving platforms undeserved legitimacy dividends, allowing them to wrap themselves in the language of IHRL even as what is required by that body of norms remains indeterminate and contested.

* Lecturer on Law & S.J.D. Candidate, Harvard Law School; Affiliate, Berkman Klein Center for Internet & Society. Many thanks to Martha Minow, Jack Goldsmith, Susan Benesch, Sam Bookman, Elena Chachko, Brenda Dvoskin, Barrie Sander, participants at the UC Irvine School of Law Symposium on Transnational Legal Ordering of Privacy and Speech, and especially to David Kaye for his tireless work in this field, inspiring leadership, and willingness to engage in good faith debate in the best tradition of freedom of expression. Huge and sincere thanks, too, to Ally Myers, Amelia Haselkorn, Sharon Baek, Jonathan Widjaja and all the student editors that shepherded this piece through the production process; their careful attention to this article significantly improved it. Any limits on this Article's correctness remain my fault alone.

INTRODUCTION

In remarkably short order, there has been growing convergence around the idea that major social media platforms should use international human rights law (IHRL) as the basis for their content moderation rules. The argument was spearheaded by David Kaye during his tenure as U.N. special rapporteur on the promotion and protection of the right to freedom of opinion and expression. Kaye made the first comprehensive case for the proposal in his mid-2018 report to the U.N. Human Rights Council.[1] The call has been increasingly echoed by academics and civil society[2] and, in a coup for the movement, platforms themselves have

1. David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 70, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018); *see also* Evelyn Douek, *U.N. Special Rapporteur's Latest* Report *on Online Content Regulation Calls for 'Human Rights by Default'*, LAWFARE (June 6, 2018, 8:00 AM), https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default.

2. *See, e.g.*, Barrie Sander, Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation, 43 FORDHAM INT'L L.J. 939, 966–69 (2020); Evelyn Mary Aswad, The Future of Freedom of Expression Online, 17 DUKE L. &

begun suggesting they do or will incorporate IHRL into their content moderation governance systems.[3]

But why have these legendarily growth-obsessed companies been so quick to voluntarily say they are jumping on this bandwagon when the advocates for this approach generally envision it operating as a constraint on their operations? In the realm of possible reasons, there are both encouraging and less encouraging answers.

For the glass-half-full types, there is the straightforward explanation that perhaps these companies do genuinely care about human rights. If so, the notion that platforms should look to IHRL in writing and enforcing the rules for what they allow on their services has intuitive appeal. On a more practical level, the major tech platforms are inherently international and generally insist on having a single global set of content standards to the extent possible.[4] Platforms' rules have significant ramifications for their users' freedom of expression, privacy, equality, and many other rights and interests that IHRL speaks to. IHRL is therefore a seemingly obvious place to turn to for their global rules affecting rights.

Although IHRL is primarily addressed to states, the Human Rights Council adopted the United Nations' Guiding Principles on Business and Human Rights (UNGPs) in 2011, which provide a framework (albeit non-binding) of expected conduct for private actors with respect to human rights.[5] The UNGPs describe a "three-pillar framework" to (1) protect and (2) respect human rights, and (3) remedy any abuses as a result of business-related activities.[6] This Article focuses on the second pillar—the corporate responsibility to "respect" human rights—because it is through constructing content moderation systems incorporating IHRL that platforms can articulate respect for human rights in the course of determining exactly what IHRL would require in any context. Of course, this merges with the

---

TECH. REV. 26, 67 (2018); Jillian C. York & Corynne McSherry, Content Moderation is Broken. Let Us Count the Ways., ELEC. FRONTIER FOUND. (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways; ELIŠKA PÍRKOVÁ & JAVIER PALLERO, TWENTY-SIX RECOMMENDATIONS ON CONTENT GOVERNANCE: A GUIDE FOR LAWMAKERS, REGULATORS, AND COMPANY POLICY MAKERS 35 (2020).

3. Jack Dorsey (@jack), TWITTER (Aug. 10, 2018, 9:58 AM), https://twitter.com/jack/status/1027962500438843397; 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/; Patrick McGee, *Apple Commits to Freedom of Speech After Criticism of China Censorship*, FIN. TIMES (Sept. 3, 2020), https://www.ft.com/content/a88f5d3d-0102-4616-8b3f-cb0661ba305d.

4. Monika Bickert, *Defining the Boundaries of Free Speech on Social Media*, *in* THE FREE SPEECH CENTURY 254, 260 (Lee C. Bollinger & Geoffrey R. Stone eds., 2018).

5. *See* Human Rights Council Res. 17/4, ¶ 1, U.N. Doc. A/HRC/RES/17/4 (July 6, 2011); John Ruggie (Special Representative of the Secretary-General), *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter *UNGPs*].

6. *UNGPs*, *supra* note 5, ¶ 6.

third pillar of "remedying" human rights abuses where platform appeals systems provide "remedies" for mistaken decisions that violate human rights.[7]

The UNGPs are non-binding but are an important tool for holding companies accountable for their human rights impacts. Their status as amongst the most widely backed global norms gives them normative and instinctive appeal for companies looking for global standards to guide their decisions with respect to rights.

But there is also a less optimistic possibility: companies are embracing the terminology so readily because they know that, in reality, it will not act as much of a constraint at all. As non-binding norms, there is no mechanism to coerce a company into compliance. Furthermore, lack of transparency, information asymmetries, and complexities in discerning the exact nature of a company's obligations make direct enforcement difficult. A variant of "bluewashing,"[8] companies can wrap themselves in the language of human rights, co-opting its legitimacy at little cost.

Indeed, it may be even worse than that: given the momentum around the call for companies to adopt IHRL standards, it would almost be irrational for companies *not* to embrace these standards if they do not require substantive changes. As more companies sign on with little sign of radical adjustments to their operations or business models, IHRL becomes signaling and mere cheap talk if it does not bring actual accountability or constraint. This is the prospect explored in this Article.

Despite this possibility, there has been almost no strong dissent from the proposition that IHRL should be adopted by companies as the basis for their rules.[9] This paper therefore aims to fill this gap with a sympathetic critique of the contributions IHRL can make to content moderation and the very real limits of IHRL as a practical guide to what platforms should do in many, if not most, difficult cases.

I say my critique is sympathetic because I share many of the commitments and aspirations of those that propound IHRL as a solution to content moderation's

---

7. The issue of actual enforcement of platform rules is a hugely important question: ideal platform policies on paper (or webpages) mean nothing if they are not consistently and accurately enforced. But for reasons of scope, this paper largely focuses on the question of what IHRL requires those rules to be in the first place.

8. Daniel Berliner & Aseem Prakash, *"Bluewashing" the Firm? Voluntary Regulations, Program Design, and Member Compliance with the United Nations Global Compact*, 43 POL'Y STUD. J. 115, 116 (2015) (defining "bluewashing" as when firms use engagement with United Nations initiatives to figuratively drape themselves in the blue UN flag in order to distract stakeholders from their real, as opposed to cosmetic, poor environmental or human rights records).

9. A recent exception is Brenda Dvoskin, *Why International Human Rights Law Cannot Replace Content Moderation*, MEDIUM (Oct. 8, 2019), https://medium.com/berkman-klein-center/why-international-human-rights-law-cannot-replace-content-moderation-d3fc8dd4344c. Others have also offered constructive criticisms. *See, e.g.*, Sander, *supra* note 2, at 968–70; Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, 38 YALE J. ON REGUL. ONLINE BULL. 86, 90 (2020).

current woes. I too search for a principled way to constrain the private power that a few dominant companies have over some of the most important channels of expression in the modern age and believe it to be one of the most pressing human rights questions today.[10] This paper does not align itself with the "archaic way of thinking"[11] that the private nature of social media companies means they have no duty to uphold the human rights of the people affected by their operations. Nevertheless, I see significant limitations on what IHRL offers in practice, and I suggest that failing to acknowledge these limitations will only serve the interests of platforms rather than their users by giving platforms undeserved legitimacy dividends. Proponents agree that IHRL is not a panacea for our current content moderation woes; here, I seek to catalogue more comprehensively why not.

I proceed as follows: Part II reviews the growing consensus for an IHRL-based approach to content moderation, and the benefits of such an approach. Part III turns to the considerable limits on what IHRL can bring to content moderation systems and argues that invoking IHRL without being attentive to these limitations could frustrate the goals of those seeking to promote human rights online. Part IV illustrates this by reference to two of the most contentious issues in content moderation: hate speech and election interference. Part V briefly suggests some next steps in the agenda for the protection of human rights in content moderation.

There is perhaps no more consequential debate for the future of free expression than how to legitimate and constrain platforms' content moderation. For IHRL to be relevant in the platform era, it needs to find purchase online and in spaces run by private companies. But this requires being clear-eyed about the enormity of that task and the significant obstacles that stand in the way of its realization.

I. WHAT INTERNATIONAL LAW CAN OFFER

As a handful of major tech companies have become ever more important "Deciders" about what can and cannot be said in some of the most important modern forums for speech,[12] there has been increasing concern about the opacity

10. *See, e.g.*, Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality & Probability*, 121 COLUM. L. REV. (forthcoming 2021).

11. DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET 119 (2019) [hereinafter KAYE, SPEECH POLICE]; David Kaye, *A New Constitution for Content Moderation*, MEDIUM (June 25, 2019), https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf [hereinafter Kaye, *A New Constitution*].

12. *See, e.g.*, Jeffrey Rosen, *The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google*, 80 FORDHAM L. REV. 1525, 1536 (2012); Marvin Ammori, *The "New" New York Times: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2273–78 (2014); Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1631–35 (2018); Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2021 (2018).

and arbitrariness with which they exercise this considerable power.[13] The decisions these companies make have enormous consequences for both individuals and societies. In the last few years alone, platforms have been implicated in rights infringements in contexts as diverse as genocide,[14] election interference,[15] widespread harassment and abuse,[16] and terrorist attacks.[17] Content moderation problems are human rights problems. (For simplicity, this paper focuses on users' freedom of expression, but many of the arguments that follow would apply equally to the entire corpus of human rights.[18])

In the early days of the internet, platforms adopted what I have elsewhere called a "posts-as-trumps" framework that generally proceeded on the assumption that platforms would by-and-large not interfere with user content.[19] This was encouraged and enabled by generally broad immunities provided by formal law.[20] But in the past few years, societal and regulatory expectations of platforms have changed markedly, and content moderation rules have become ever-more expansive legal-esque codes.[21] Platforms no longer focus only or primarily on the speech rights of posters, but are taking a broader (albeit still too limited) view of their responsibility and recognizing other societal interests as well.[22] But as platforms draw these lines, pressing questions arise as to how their decisions can be made principled and legitimate. No pre-existing body of rules easily applies.

---

13. *See, e.g.*, TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 5–6 (2018); NICOLAS P. SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 28 (2019); EMILY B. LAIDLAW, REGULATING SPEECH IN CYBERSPACE: GATEKEEPERS, HUMAN RIGHTS AND CORPORATE RESPONSIBILITY 107 (2015).

14. 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; Evelyn Douek, *Facebook's Role in the Genocide in Myanmar: New Reporting Complicates the Narrative*, LAWFARE (Oct. 22, 2018, 9:01 AM), https://www.lawfareblog.com/facebooks-role-genocide-myanmar-new-reporting-complicates-narrative.

15. RENEE DIRESTA, KRIS SHAFFER, BECKY RUPPEL, DAVID SULLIVAN, ROBERT MATNEY, RYAN FOX, JONATHAN ALBRIGHT & BEN JOHNSON, THE TACTICS & TROPES OF THE INTERNET RESEARCH AGENCY 101 (2019).

16. Danielle Keats Citron & Mary Anne Franks, *The Internet as a Speech Machine and Other Myths Confounding Section 230 Reform*, U. CHI. LEGAL F. (2020).

17. Charlie Warzel, *The New Zealand Massacre Was Made to Go Viral*, N.Y. TIMES (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/opinion/new-zealand-shooting.html.

18. I also focus on purely private regulation and exclude questions of state pressure or co-optation (what Balkin has described as "collateral censorship") that would more properly be described as state action. *See* Balkin, *supra* note 12, at 116 ("Collateral censorship occurs when the state targets entity A to control the speech of another entity, B.").

19. Douek, *supra* note 10; Jonathan Zittrain, *Three Eras of Digital Governance* (Sept. 15, 2019), https://www.ssrn.com/abstract=3458435.

20. Anupam Chander, *How Law Made Silicon Valley*, 63 EMORY L.J. 639 (2014).

21. *See generally* Klonick, *supra* note 12.

22. Douek, *supra* note 10, at 24.

Unlike other private interactions with human rights, platforms are often cast in the role of adjudicating conflicts between rights of users or society. But unlike state actors, private platforms are businesses that have legitimate interests in designing and curating their services in the way they choose. Nevertheless, given their profound systemic importance, and their own claims to be akin to "public squares,"[23] users and regulators have come to expect that platforms will have regard for the public interest to at least some degree in the way they construct and enforce their rules. But according to what principles? For a long time, if human rights were considered at all, many considered that internet intermediaries were not only upholding but *expanding* human rights almost *simply by existing*, by facilitating the ever-greater flow of expression and information and creating a seemingly borderless world.[24] This view has waned, but exactly what should replace it remains fiercely contested.

Against this background, in April 2018, Kaye proposed a framework for content moderation that "puts human rights at the very centre."[25] Kaye's central argument was that "[c]ompanies 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."[26] This approach, which he called "human rights by default," would require platforms to explain their rules with greater clarity and specificity, and show that infringements on freedom of expression were narrowly tailored.

The framework in the report—which, like this Article, focused primarily on freedom of expression issues—is that provided by Article 19 of the International Covenant on Civil and Political Rights.[27] Article 19 sets out the right to freedom of expression[28] and then immediately acknowledges that the right can be subject to certain restrictions.[29] The onus is on the authority restricting speech, and a limitation must meet three conditions:[30]

---

23. Mark Zuckerberg, *A Privacy-Focused Vision for Social Networking*, FACEBOOK (Mar. 6, 2019), https://www.facebook.com/notes/mark-zuckerberg/a-privacy-focused-vision-for-social-networking/10156700570096634/; *Twitter: Transparency and Accountability: Hearing Before the H. Comm. on Energy & Com.*, 115th Cong. (2018) (statement of Jack Dorsey, CEO, Twitter, Inc.), https://docs.house.gov/meetings/IF/IF00/20180905/108642/HHRG-115-IF00-Wstate-DorseyJ-20180905.pdf [hereinafter *Twitter*].

24. Agnès Callamard, *The Human Rights Obligations of Non-State Actors*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 191, 204–05 (Rikke Frank Jørgensen ed., 2019).

25. Kaye, *supra* note 1, ¶ 2.

26. *Id.* ¶ 45.

27. G.A. Res. 217 (III) A, Universal Declaration of Human Rights, art. 19 (Dec. 10, 1948) art. 19, *adopted* Dec. 19, Dec. 16, 1966, S. Exec., 999 U.N.T.S. 171 [hereinafter, ICCPR].

28. ICCPR, *supra* note 27, art. 19(2).

29. ICCPR, *supra* note 27, art. 19(3), at 178.

30. David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 6 U.N. Doc. A/74/48050 (Oct. 9, 2019).

1. *Legality*: permissible restrictions are only those that are "provided by law"—that is, a rule restricting speech must be specified precisely, publicly, and transparently;
2. *Legitimacy*: the restriction needs to be for the purpose of one of a defined set of interests: to protect rights or reputations of others, national security, public order, public health, or morals;
3. *Necessity and proportionality*: the restriction must be necessary and the least restrictive means to achieve the purported aim.

In addition, Article 20 requires the prohibition of propaganda for war and "advocacy of national, racial or religious hatred."

There would be a number of benefits to adopting these rules as the basis for platforms' content moderation rules. In what follows, I survey these benefits in a somewhat abridged form, not least because advocates for IHRL in content moderation have made the arguments much more forcefully and eloquently elsewhere.

Here, I catalogue six main benefits: (A) using IHRL could give content moderation a legitimacy that it currently lacks; (B) as a set of global norms, IHRL aligns with major platforms' and open internet advocates' desire to have a single set of global rules, to the extent possible; (C) IHRL provides a common vocabulary for participants in debates around appropriate rules for online speech, including internal stakeholders; (D) IHRL would provide companies with a normative basis for denying authoritarian requests to censor content on their services; (E) IHRL encompasses a set of procedural norms, as well as substantive ones, which can help constrain platforms' arbitrary exercise of power; and (F) IHRL is the least-worst option available.

A. Legitimacy

As platforms are begrudgingly drawn ever further into the task of transparently writing and justifying rules for how they will deal with content on their services, their apparent arbitrariness and lack of consistent enforcement has caused a severe legitimacy deficit.[31] For years, academics and civil society have been pointing out the awesome and arbitrarily exercised power of these companies,[32] but the past few years have seen these views break through into mainstream consciousness and cause

31. Evelyn Douek, *Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility*, 21 N.C. J.L. & TECH. 1, 18 (2019).

32. I cannot possibly do justice to all the important early voices here, but for some notable examples, see for example, REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM (2012); Jillian C. York, *Policing Content in the Quasi-Public Sphere*, OPENNET INITIATIVE 1 (2010), https://opennet.net/policing-content-quasi-public-sphere.

a societal "techlash."[33] The delegitimization of existing systems of governance and the frustration with the ad hoc way platforms write and enforce rules has been so widespread that even Facebook CEO Mark Zuckerberg has been forced to concede that "[l]awmakers often tell me we have too much power over speech, and frankly I agree. . . . [W]e need a more standardized approach."[34]

On its face, IHRL seems to remedy this legitimacy deficit. As a set of global norms based on as close to universal state consent as anything, IHRL offers "a sense of global legitimacy, credibility, and appeal, especially outside the United States."[35] In the words of the UN High Commissioner for Human Rights, IHRL is "the only tested international set of rules and principles resulting from decades of debate and collaboration among States and international legal experts from around the globe."[36] As Kaye explains, "it would offer a globally recognized framework for designing [the] tools [to accommodate varied interests] and a common vocabulary for explaining their nature, purpose and application to users and State."[37]

In contrast to the current paradigm of platforms largely just "making rules up,"[38] IHRL ostensibly offers a form of constraint on decision-making as a set of established norms external to the platforms.

B. Global Rules

The "global" nature of IHRL is an inherent part of its appeal. For many, "[t]he doctrine of human rights has aspired from the outset to be universal, to be a doctrine that applies everywhere to everyone, irrespective of nationality, culture, tradition, ideology, or social conditions."[39] The alternative is a cacophony of differing and conflicting national, regional, and even local standards.

---

33. Eve Smith, *The Techlash Against Amazon, Facebook and Google—and What They Can Do*, ECONOMIST (Jan. 20, 2018), https://www.economist.com/briefing/2018/01/20/the-techlash-against-amazon-facebook-and-google-and-what-they-can-do; Rana Foroohar, *Year in a Word: Techlash*, FIN. TIMES (Dec. 16, 2018, 5:00 PM), https://www.ft.com/content/76578fba-fca1-11e8-ac00-57a2a826423e.

34. Mark Zuckerberg, Opinion, *Mark Zuckerberg: The Internet Needs New Rules. Let's Start in These Four Areas.*, WASH. POST (Mar. 30, 2019), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.

35. Sejal Parmar, *Facebook's Oversight Board: A Meaningful Turn Towards International Human Rights Standards?*, JUST SECURITY (May 20, 2020), https://www.justsecurity.org/70234/facebooks-oversight-board-a-meaningful-turn-towards-international-human-rights-standards.

36. Letter from the U.N. High Commissioner for Human Rights to the President of the European Commission (Sept. 7, 2020), https://europe.ohchr.org/EN/Stories/Documents/2020%2009%2007%20Letter%20HC%20to%20EC%20President.pdf.

37. Kaye, *supra* note 1, at 15.

38. MATTHIAS C. KETTEMANN & WOLFGANG SCHULZ, SETTING RULES FOR 2.7 BILLION: A (FIRST) LOOK INTO FACEBOOK'S NORM-MAKING SYSTEM: RESULTS OF A PILOT STUDY 28 (2020).

39. Antonio Cassese, *A Plea for a Global Community Grounded in a Core of Human Rights*, *in* REALIZING UTOPIA 136, 136 (Antonio Cassese ed., 2012).

IHRL's aspiration for universality aligns with that of platforms. Platforms generally prefer, for reasons of ease and product experience, to have a single set of global rules. As Facebook's head of policy has written, "[t]he borderless and dynamic nature of social media communications requires standards that are globally applied."[40] Adopting IHRL as this set of global norms avoids a "race to the bottom" caused by adopting the rules of any particular jurisdiction. In some ways, this also keeps alive the optimism and hope of a borderless internet. As Goldsmith and Wu summarized the view of "internationalists" advocating for international norms in the early days of the internet, "[n]ot only would internationalism solve the problem of conflicting laws, it also offered the promise of better laws. . . . An international approach could not only clear up confusion and conflict, but it could also wash clean the prejudice and ignorance hiding in the basement of national government."[41]

Therefore, ambitions for IHRL as a universalizing body of law finds common ground with those of platforms and internet-optimists in the use of IHRL as a basis for a system of global governance of online speech.

C. Common Vocabulary

Because the substantive rules for platform standards will almost always be a matter of reasonable disagreement and contestation, legitimacy for content moderation systems needs to be earned not by unilaterally issuing "correct" rules but by channeling debate about what those rules should be.[42] In fulfilling this need, IHLR can provide "a common conceptual language" for explaining, justifying, and challenging content moderation decisions.[43] Even when IHRL does not dictate a specific substantive outcome, it can provide a framework and vocabulary for argumentation,[44] which can facilitate a process of deliberation and public reason-giving that endows decisions with greater legitimacy.[45] Engaging in reasoning and justifying decisions is a way of showing respect for and dignifying the interests of those affected by decisions,[46] as well as providing accountability through public and transparent reasoning.[47] (This is, to many, a weak form of accountability to be sure,

40. Bickert, *supra* note 4, at 260. It is fascinating to observe that in dismissing the appropriateness of a single set of laws as being the basis of those standards, Bickert did not even consider IHRL when writing in late 2018—so rapid has the movement, spearheaded by Kaye, progressed.

41. JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET?: ILLUSIONS OF A BORDERLESS WORLD 26–27 (2006).

42. *See* Douek, *supra* note 31, at 66–76.

43. Sander, *supra* note 2, at 967.

44. *Id.* at 968.

45. On the way in which frameworks and argumentation can provide this form of legitimacy, see, for example, Mattias Kumm, *The Idea of Socratic Contestation and the Right to Justification: The Point of Rights-Based Proportionality Review*, 4 L. & ETHICS HUM. RTS. 141 (2010).

46. Monica Hakimi, *Why Should We Care About International Law?*, MICH. L. REV. 1283, 1302 (2020); *see* Douek, *supra* note 10, at 21–22.

47. Hakimi, *supra* note 46, at 1305; *see* Douek, *supra* note 31, at 66–76.

but it remains a valuable one.) I argue below that global consensus not only does not exist but likely never will.[48] But the benefit of a common vocabulary does not depend on substantive agreement; indeed, its virtue lies in providing the framework and vocabulary for ongoing and unavoidable debates.

This also applies *within* companies, where IHRL can give *internal* stakeholders the language and normative backing to articulate concerns about and potential constraints on company decision-making.[49] A more IHRL-centered discourse around the impacts of content moderation decisions provides a coherent framework and greater normative force to conversations at all stages of platform operations and can operate as a positive feedback loop.

D. Stiffened Spines

The normative force of IHRL is at its greatest and most important in stiffening the spines of companies against obvious state abuses of human rights. Adopting IHRL can help companies stand up to governmental demands that companies infringe users' rights. As Kaye argued, "[a] human rights framework enables forceful normative responses against undue State restrictions—provided companies play by similar rules."[50] When governments demand censorship,

> [i]t is much less convincing to say to authoritarians, "We cannot take down that content because that would be inconsistent with our rules," than it is to say, "Taking down that content would be inconsistent with the international human rights our users enjoy and to which your government is obligated to uphold."[51]

Therefore, as IHRL promises global substantive standards, it also provides normative backing for denying governments who would have platforms deviate from those standards. This was exactly how Facebook invoked IHRL when it said it would challenge a legal order from the Thai government to block access to a group with one million members; the group criticised the country's king, which is illegal under the country's *lèse majesté* laws.[52]

This is a compelling argument and an important contribution that IHRL can make. The value of this role is critical given the rapid rise of digital authoritarianism around the world, as repressive regimes seek to crack down on fundamental

48. *See infra* Part III(A)–(C).

49. I am grateful to Dierdre Mulligan and Jen Daskal for stressing the importance of this point.

50. Kaye, *supra* note 1, at 14; *see also* KAYE, SPEECH POLICE, *supra* note 11, at 18.

51. Kaye, *A New Constitution*, *supra* note 11.

52. Patpicha Tanakasempipat, *Facebook Blocks Group Critical of Thai Monarchy Amid Government Pressure*, REUTERS (Aug. 24, 2020, 10:44 AM), https://www.reuters.com/article/us-thailand-facebook/facebook-blocks-group-of-one-million-critical-of-thai-monarchy-amid-government-pressure-idUSKBN25K25C.

freedoms online.[53] IHRL's role in these rainy-day scenarios is vital. But, as I explore below, the question of whether a state is violating its obligations under IHRL is often a much easier one to answer (and still by no means determinate) than how a private actor should directly apply IHRL in making its decisions.[54]

E. Process

IHRL should not be understood merely as a set of substantive rules to be picked up, plugged into community standards, and applied by platforms. A large, if not dominant, part of the benefits offered by IHRL to content moderation is the procedural obligations it requires. The idea is not that platforms look to existing determinations of how restrictions on freedom of expression comply with the requirements of legality, legitimacy, necessity, and proportionality, but that they must meet each of these requirements *themselves* in formulating their rules.

This is a fundamental point for two reasons. First, as I explore further below,[55] the substantive rules in IHRL are not settled and, perhaps more importantly, leave many gaps that need to be filled in practice. Second, it is the *process* of reasoning that redounds to the legitimacy benefits discussed above. It is not the mere fact that interests have been recognized and balanced in the formulation of a rule, but the explanation of how the decision-maker has done so that makes a decision more likely to be deemed worthy of respect.[56]

The application of IHRL's procedural obligations is the easiest to apply to company activities, and activists have long used IHRL to emphasize company obligations of transparency, due process, and remediation.[57] Kaye similarly drew attention to IHRL's procedural requirements as a necessary part of fully empowering users and enhancing accountability, especially transparency[58] and remediation, for people wronged by platform decisions.[59]

It is worth noting, however, that there is a tendency to oversimplify the guidance that IHRL offers and its applicability to content moderation ecosystems. The general understanding is that while there may be variation or ambiguity around

---

53. ADRIAN SHAHBAZ & ALLIE FUNK, FREEDOM ON THE NET 2019: THE CRISIS OF SOCIAL MEDIA 1 (2019).

54. I will return to this in Parts III(A)–(C) below.

55. *See infra* Part III(A)–(C).

56. BEN BRADFORD, FLORIAN GRISEL, TRACEY L. MEARES, EMILY OWENS, BARON L. PINEDA, JACOB N. SHAPIRO, TOM R. TYLER & DANIELI EVANS PETERMAN, REPORT OF THE FACEBOOK DATA TRANSPARENCY ADVISORY GROUP 34 (2019); *see also* Tom R. Tyler, *Procedural Justice, Legitimacy, and the Effective Rule of Law*, 30 Crime & Just. 283 (2003); Douek, *supra* note 31, at 67.

57. Molly K. Land, *Regulating Private Harms Online: Content Regulation Under Human Rights Law*, *in* HUMAN RIGHTS IN THE AGE OF PLATFORMS 285, 288 (Rikke Frank Jørgensen ed., 2019).

58. Kaye, *supra* note 1, at 20 ("The companies must embark on radically different approaches to transparency at all stages of their operations, from rule-making to implementation and development of "case law" framing the interpretation of private rules.").

59. *Id.* at 18.

substantive protections, "there is more widespread agreement regarding the procedures that are essential to ensure protection for the categories of speech that are deemed legal."[60] These requirements should include an individualized determination by an independent arbiter on a case-by-case basis.[61]

But the scale of content moderation makes such individualized assessment, even if only for appeals of initial decisions, likely impractical and impossible.[62] To deny such an assessment will not necessarily always be an infringement on rights: due process rights—what process is due—have always been highly contextual and systemic.[63] The extent of individualized assessment required will turn on many factors, including the particular category of content in question, the average accuracy of initial decisions, the severity of the consequences of the initial decision, and a cost-benefit assessment of the value offered by any additional procedure.

It is beyond my present scope to offer a full account of these factors, but it is sufficient to note that even in the realm of procedural requirements, IHRL does not necessarily provide a determinate answer of what is required. Furthermore, as I explore further below, as valuable as procedural rules are, they still need to be backed by substantive commitments to be made meaningful.[64]

F. The Least-Worst Option

Finally, and the importance of this should not be underestimated, IHRL is appealing because there is no obvious and compelling alternative. Designers of content moderation systems are in search of a way to legitimize global rules about human rights written and enacted by extremely powerful private companies. As Bowers and Zittrain observe, "[t]he inwards-looking, largely public relations-oriented content governance models so widely deployed today are unsatisfying."[65] They are largely untethered from any particular normative commitments, leaving them unconstrained and arbitrary. The question of whether IHRL is a good basis for content moderation rules should be asked not in a vacuum but as *compared to what?* IHRL need not be perfect to be the least-worst option. With current systems delegitimized and the laws of individual countries inappropriate for international application, IHRL has a strong claim to being the worst option except for all the rest.

---

60. Dawn C. Nunziato, *The Beginning of the End of Internet Freedom*, 45 GEO. J. INT'L L. 383, 396 (2014).

61. Emma J. Llansó, *No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem*, BIG DATA & SOC'Y 1, 4 (2020); *see generally* Nunziato, *supra* note 60.

62. *See* Evelyn Douek, *Verified Accountability: Self-Regulation of Content Moderation as an Answer to the Special Problems of Speech Regulation*, HOOVER INST. 1, 8–11 (2019).

63. *Id.* at 9.

64. *See infra* Part III(D).

65. John Bowers & Jonathan Zittrain, *Answering Impossible Questions: Content Governance in an Age of Disinformation*, 1 HARV. KENNEDY SCH. MISINFO. REV. 1, 5 (2020).

II. THE LIMITS OF INTERNATIONAL LAW IN CONTENT MODERATION

The movement for IHRL to be adopted as the basis for content moderation rules has garnered considerable momentum because of these strong arguments in its favor. For a long time, platforms argued that "human rights are neither their concern nor their responsibility"[66] and were "reluctant to view content moderation undertaken to enforce their terms of service (TOS) as a human rights issue."[67] It is therefore no small feat to have them recognize otherwise.

But to return to the question asked in the introduction: *why* have platforms suddenly voluntarily signed on? The explanation that they genuinely care about human rights is not enough: if it were, there are many other steps they could take including, at a minimum, being much more transparent about their services, hiring more content moderators (and providing them better working conditions), or introducing more friction and reducing the extent to which their algorithms optimize for engagement. (And, notably, on the rare occasions platforms have taken such steps, they rarely articulate them as being responsive to IHRL obligations.) There are very real limits on what companies are prepared to voluntarily do in the name of human rights, and yet they have relatively quickly been prepared to profess to adopt IHRL in their content moderation rules.[68]

This Part argues that companies are prepared to make these commitments largely because simply adopting IHRL as the basis of content moderation would not constrain the operations of these platforms to any significant extent. As a result, it would likely not give them the legitimacy benefits they hope for; but it may still get them more than they have earned if their adoption of IHRL is not viewed critically.

This is not to dismiss the contributions that IHRL can make in this area. But this requires being very upfront about its limitations so that its weaknesses can be addressed. These extend well beyond the most simple and obvious: their non-binding nature. The inability to directly enforce IHRL in this context is in part the product of, and at the very least exacerbates, other weaknesses. These weaknesses include the following: (A) the consensus with respect to IHRL is not simple or universal and in fact is highly contested, especially when it comes to freedom of expression; (B) IHRL is not a single, self-contained, cohesive body of rules but includes gaps, inconsistencies, and is subject to differing interpretations; (C) there is a large degree of indeterminacy in IHRL norms that would leave platforms with extensive discretion in many if not most hard cases; (D) given this, there is room for platforms to co-opt the language and legitimacy of IHRL; (E) even if platforms were to try to adopt IHRL in good faith, they lack the information or competency to conduct the assessment and balancing of interests necessary; (F) the

66. MACKINNON, *supra* note 32, at 273.

67. DAVID SULLIVAN, ASS'N FOR PROGRESSIVE COMM., BUSINESS AND DIGITAL RIGHTS: TAKING STOCK OF THE UN GUIDING PRINCIPLES FOR BUSINESS AND HUMAN RIGHTS IN THE ICT SECTOR 16 (2016).

68. *See supra* Part I.

indeterminacy and competency critiques apply equally to the procedural norms under IHRL as they do to the substantive norms; and (G) the platform era requires a paradigm-shift in thinking about rights, from individualistic to systemic, and IHRL has not yet developed jurisprudence or tools to deal with this fundamental change.

It is worth noting at the outset that many people have suggested that IHRL obligations should vary depending on the size and systemic importance of the platform in question.[69] This Article does not consider this point, except to note that it is only further indicative of the open nature of many of the relevant questions. My present focus, however, is on the indeterminacy that remains even if the analysis is confined to the largest platforms that claim to be "public squares."[70]

A. Highly Contested

Even if it could be said that there is a single cohesive body of law called "IHRL" (I contest this in the following sub-parts), the unanimity with respect to the legitimacy of this body of norms is far from complete. As Anthea Roberts has comprehensively documented, "international lawyers' romantic understanding of themselves and their field as universal and cosmopolitan" is incomplete and sometimes misleading.[71] Far from being a cohesive whole, international law is characterized by different understandings from different relevant communities, while certain communities dominate in shaping the field.[72] To the extent that there is universality, this itself reflects a set of values of cosmopolitanism and belief in the efficiency of uniformity that is contested.

There is no area in which this is more true than with respect to freedom of expression. There are still significant numbers of states with reservations to the freedom of expression provisions in international treaties.[73] As one judge of the International Criminal Tribunal for Rwanda commented, "[t]he number and extent of the reservations [to these provisions] reveal that profound disagreement persists in the international community[;] . . . a consensus among states has not crystallized."[74] As even Evelyn Aswad—a staunch advocate for IHRL in content moderation—acknowledges, "[t]he scope of ICCPR Article 20 remains controversial to this day."[75] Famously, of course, this includes the United States, and many of those accustomed to First Amendment jurisprudence continue to

69. *See, e.g.*, Land, *supra* note 57; KATE JONES, CHATHAM HOUSE, ONLINE DISINFORMATION AND POLITICAL DISCOURSE: APPLYING A HUMAN RIGHTS FRAMEWORK 6 (2019).

70. Zuckerberg, *supra* note 23; *Twitter*, *supra* note 23.

71. ANTHEA ROBERTS, IS INTERNATIONAL LAW INTERNATIONAL? 6 (2017).

72. *See, generally*, ROBERTS, *supra* note 71.

73. Amal Clooney & Philippa Webb, *The Right to Insult in International Law*, 48 COLUM. HUM. RTS. L. REV. 1, 20 (2017).

74. Nahimana v. Prosecutor, Case No. ICTR-99-52-A, Judgment on Appeal, at 376 (Nov. 28, 2007) (Meron, J., dissenting).

75. Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 37 (2018).

reject any notion of convergence towards the international standards.[76] My point here is not to take a position in this debate, but simply to note that under these conditions it cannot truly be said that IHRL reflects a global consensus.

Advocates argue that just because states fail to live up to their obligations does not mean that the obligations themselves are indeterminate or contested.[77] Aspirational norms can still have value, to be sure. But there is a tension in relying on state consent as a basis for IHRL's legitimacy and ignoring state divergence or contestation on difficult edge cases.

Here, and in what follows, it is important not to overstate my argument or the extent of the divergence. There are "many areas in which international and regional human rights law have substantively converged with the US protection of freedom of expression."[78] Nevertheless, focusing on the marginal and exceptional cases where this is not true is not misplaced because it is precisely in those cases that IHRL has the most work to do. Where there *is* convergence and unanimity, IHRL does not in substance add much beyond rhetorical legitimation. Rules that match both universal and local norms are easy cases. Facebook does not—or, at least, should not—need IHRL to know that incitement to genocide is wrong. But it is in the harder cases where reasonable minds can differ that platforms are most in need of constraint and legitimation. And yet it is in those cases—the very cases where IHRL stands to offer the most—where there is a divergence of approach among states and IHRL's indeterminacy are also the greatest.

B. No Global Norms

As I have written elsewhere:

> It is something of a misnomer to speak of international human rights law as if it is a single, self-contained and cohesive body of rules. Instead, these laws are found in a variety of international and regional treaties that are subject to differing interpretations by states that are parties to the conventions as well as international tribunals applying the laws.[79]

In the context of a right to insult, for example, Amal Clooney and Philippa Webb have shown IHRL is sometimes confusing or inconsistent and treaties have not been applied clearly or consistently.[80] This applies to the Human Rights Committee's interpretations of the ICCPR and other international treaties, and is

76. *See, e.g.*, Suzanne Nossel, *The American Approach to Free Speech Is Flawed—But It's the Best Option We Have*, SLATE, (July 28, 2020, 9:00 AM), https://slate.com/technology/2020/07/america-free-speech-first-amendment-misinformation.html; Clooney & Webb, *supra* note 73.

77. Evelyn Aswad, *To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?*, 77 WASH. & LEE L. REV. 609, 632–33 (2020).

78. Sarah H. Cleveland, *Hate Speech at Home and Abroad*, *in* THE FREE SPEECH CENTURY 210, 210 (Lee C. Bollinger & Geoffrey R. Stone eds., 2019).

79. Douek, *supra* note 1.

80. Clooney & Webb, *supra* note 73, at 36.

compounded by confusion and contradictions in various regional instruments and their interpretation by regional bodies.[81] As Barrie Sander notes, while platforms can look to "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."[82] Indeed, "[g]lobal human rights norms—from binding treaty provisions to soft law recommendations of international human rights bodies—are diverse, nuanced, evolving, sometimes inconsistent, and contested, *especially* in the area of freedom of expression."[83] In short, as Benesch summarizes, "human rights law on speech is confusing and not always applicable to private companies."[84]

Aswad has argued that the UNGP's should be taken to require companies to align their rules with "*international* human rights law," and "*regional* human rights instruments (and monitoring bodies) are not *international* human rights instruments (and monitoring bodies)."[85] But regional jurisprudence should not be so quickly dismissed. Aside from the fact that regional treaties are definitionally part of international law, they also play an important normative and practical role in IHRL jurisprudence and norm-setting. The language of "IHRL" narrowly conceived as only the ICCPR remains broad and general, leaving many gaps to be at best filled, and at worst exploited. As Clooney and Webb conclude, "[i]nternational law on free speech is . . . not well defined in the jurisprudence of the U.N. bodies, leaving far too much scope for abuse. The fears voiced by many states during the drafting of the ICCPR and CERD regarding free speech have turned out to be well-founded: the terms used are too vague and susceptible to abuse."[86] Kaye himself has argued that the jurisprudence of regional bodies plays an important role in IHRL:

> [S]ome might say that human rights law is too general for the companies to apply. But companies have ample jurisprudence to draw from. . . . This jurisprudence can be found in the European Court of Human Rights, the Inter-American Court for Human Rights, the emerging jurisprudence of regional and sub-regional courts in Africa, national courts in democratic societies, treaty bodies that monitor compliance with their norms, and the work of UN and regional human rights mechanisms. It is not an answer to say the law does not exist because some look down upon it as a lesser form of law, or because of ignorance that this body of law even exists.[87]

---

81. *See also* Parmar, *supra* note 35.

82. Sander, *supra* note 2, at 40.

83. Parmar, *supra* note 35 (emphasis added).

84. SUSAN BENESCH, PROPOSALS FOR IMPROVED REGULATION OF HARMFUL ONLINE CONTENT 15 (2020).

85. Aswad, *supra* note 75, at 44; *see also* Aswad, *supra* note 77, at 642.

86. Clooney & Webb, *supra* note 73, at 44.

87. Kaye, *A New Constitution*, *supra* note 11.

Looking to these norms is not only practically necessary to fill in relevant gaps, but also normatively desirable: to the extent that speech decisions are highly contextual and need to take account of local norms and facts, the jurisprudence of regional bodies will obviously be relevant and enlightening.[88]

International lawyers "often resist emphasizing national or regional approaches because they are seen as potentially threatening to the field's universalist aspirations."[89] But it is a different set of universalist aspirations that have been partially responsible for some of the greatest failures in content moderation and early internet governance. Jack Goldsmith and Tim Wu proved prescient in their observation that alongside globalization would remain "the determined preservation of difference, the deliberate resistance to homogenizing influence" as different places fight to retain their own cultures and values.[90] They defended this as normatively desirable: "[T]here is very little to say in favor of a single global rule for Internet speech. . . . These dramatically different attitudes toward proper speech among the mature democracies reflect important differences among the peoples that populate these countries—differences in culture, history, and tastes."[91]

Platforms' unduly unidimensional understanding of free expression, and failure to adjust to local contexts, is one of the most naïve assumptions of the early platform era. It should not be repeated in the adoption of IHRL in content moderation. Indeed, IHRL is a body of norms that welcomes diversity: this is a strength in many respects but a weakness when looking for a determinate source of constraint on platform decisions.

Even if, as I have just argued against, universal norms are the ultimate goal, interactions between regional systems and the universal one can lead to "converging norms and procedures in an overarching interdependent and dynamic system. In many respects they are thinking globally and acting regionally."[92] Abandoning regional differentiation for top-down imposition of norms is unlikely to be successful.

"Thinking globally and acting regionally" is, in many ways, exactly what critics have been calling on social media platforms to do. One of the main criticisms that global platforms have encountered in recent years has been their lack of attention to different demands of varying contexts.[93] Regional bodies are an important way

88. JONES, *supra* note 69, at 31.

89. ROBERTS, *supra* note 71, at 20.

90. GOLDSMITH & WU, *supra* note 41, at 183.

91. *Id.* at 150.

92. Dinah Shelton, *The Promise of Regional Human Rights Systems*, *in* THE FUTURE OF INTERNATIONAL HUMAN RIGHTS 351, 356 (Burns H. Weston & Stephen P. Marks eds., 1999).

93. *See, e.g.*, SIVA VAIDHYANATHAN, ANTISOCIAL MEDIA: HOW FACEBOOK DISCONNECTS US AND UNDERMINES DEMOCRACY 27 (2018); Chinmayi Arun, *Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y 2-3 (2018); KAYE, SPEECH POLICE, *supra* note 11, at 117.

of understanding local context. It may be that *regional* interpretations are different because the context is different. That is: while regional interpretations may not be invoked to justify departure from decisions under the ICCPR,[94] different *outcomes* may not be the result of different rules, but instead a product of different circumstances.

Malcolm D. Evans has even argued that the emergence of "variegated and complex pattern of jurisdictional networks" in IHRL is a sign of its growing maturity.[95] Dinah Shelton agrees: "Each uses the jurisprudence of the other systems and amends and strengthens its procedures with reference to the experience of the others. In general, their mutual influence is highly progressive, both in normative development and institutional reform."[96]

All of these caveats about the diversity of approaches in various bodies interpreting rights apply especially forcefully in the context of the novel and open questions that content moderation raises. Brenda Dvoskin has examined one useful example: the question of whether the use of filtering systems at the time of upload to detect and block certain kinds of content (that is, a form of prior restraint) is consistent with international law. The Court of Justice of the European Union has held that in certain cases such artificial intelligence filters are not only consistent with European law but may in fact be required by it.[97] But, as Dvoskin observed, "the remedies the CJEU offered are explicitly forbidden in the American Convention on Human Rights. Specifically, the American Convention forbids prior restraint of speech and does not allow for the type of balancing test that is permissible in the European context."[98]

There is no decision as yet interpreting ICCPR Article 19's application to upload filters specifically, or authoritative guidance about such interpretation, save reports of experts such as Kaye. There may be a case or comment in time, but in the meantime, content moderation systems will make billions if not trillions of decisions and any slow-moving jurisprudence will likely be almost immediately

94. Kaye, *supra* note 30, at 9–10.

95. Malcolm D. Evans, *The Future(s) of Regional Courts on Human Rights*, *in* REALIZING UTOPIA 261, 273 (Antonio Cassese ed., 2012).

96. Shelton, *supra* note 92, at 356.

97. Case C-18/18, Glawischnig-Piesczek v. Facebook Ir., ECLI:EU:C:2019:821, ¶¶ 45-46 (Oct. 3, 2019).

98. Brenda Dvoskin, *Why International Human Rights Law Cannot Replace Content Moderation*, MEDIUM (Oct. 8, 2019), https://medium.com/berkman-klein-center/why-international-human-rights-law-cannot-replace-content-moderation-d3fc8dd4344c. This was the opinion of the former Special Rapporteur for Freedom of Expression of the Inter-American Convention Catalina Botero, Catalina Botero Marino (Special Rapporteur for Freedom of Expression), *Freedom of Expression and the Internet*, Inter-Am. Comm'n H.R., at ¶ 88, OEA/Ser.L/V/II CIDH/RELE/INF. 11/13 (Dec. 31, 2013), http://www.oas.org/en/iachr/expression/docs/reports/2014_04_08_internet_eng%20_web.pdf, and a decision of the Supreme Court of Argentina interpreting the Convention, Corte Suprema de Justicia de la Nación [CSJN][National Supreme Court of Justice], 29/10/2014, "Rodriguez, Maria Belén c/ Google Inc. s/ daños y perjuicios," Fallos (2014-R-522) (Arg.), http://www.saij.gob.ar/corte-suprema-justicia-nacion-federal-ciudad-autonoma-buenos-aires-rodriguez-maria-belen-google-inc-otro-danos-perjuicios-fa14000161-2014-10-28/123456789-161-0004-1ots-eupmocsollaf.

superseded by advances in technology. As platforms continue to operate, if they want to apply IHRL, they themselves will need to make a number of choices about what that body of law requires.

Again, while inconsistencies and gaps in these bodies of law cannot be dismissed, the problems they cause should also not be overstated. No body of law is completely comprehensive or consistent. Uncertainty and ambiguity are part of the evolution of any legal doctrine. But the critique that there are no global norms remains important for two reasons.

First, *no* legal system has good answers for the most difficult questions in content moderation yet. Many of these questions are still essentially open, leaving platforms with practically no guidance. International law's particularly broad language and slow-moving machinery makes this an especially acute issue.

Second, and perhaps more importantly, not acknowledging gaps or inconsistencies exacerbates the extent to which vague language can be open to abuse by denying how much agency and choice there is for private actors who assert to be simply applying IHRL. Furthermore, it risks leaving the project of advancing IHRL open to the critique that it is simply "looking out over a crowd and picking out [its] friends."[99] IHRL includes regional jurisprudence and acknowledging both the strengths and weaknesses of this is the most candid way to advance the project, as well as an attempt to avoid its manipulation and misuse.

Perhaps most importantly: even *if* a universal system that disregards regional jurisprudence is the preferred solution, it does not naturally follow that platforms themselves are the best actors to resolve inconsistencies or plug (the many) gaps.[100]

C. Indeterminacy

The question of whether basing content moderation rules on IHRL would be beneficial needs to be answered, at least in part, by examining the extent to which standards would be *different* from what they would be without looking to IHRL. And this question, in turn, depends on the extent to which IHRL would constrain platform decision-making. If IHRL still leaves platforms with wide discretion to do what they would have done anyway, its utility is significantly diminished (but not entirely because, aside from substantive rules, there are still the potential discursive and argumentative benefits[101]). The question can be posed as: Does consulting IHRL make a putative platform lawyer feel compelled to adopt a policy they would not have otherwise adopted?

The variance and flexibility in state laws about speech are enough to show considerable residual discretion in the interpretation and application of IHRL. As

99. Adam Liptak, *U.S. Court Is Now Guiding Fewer Nations*, N.Y. TIMES (Sept. 17, 2008), https://www.nytimes.com/2008/09/18/us/18legal.html (quoting Chief Justice Roberts).

100. *See infra* Part III(E).

101. *See infra* Part III(D).

Andrew Legg observes (following James Crawford), although the Human Rights Committee has largely declined to explicitly talk of giving states a "margin of appreciation" in the implementation of their obligations under IHRL, there is ample evidence supporting the proposition that it forms part of the Committee's practice and that the Committee applies limitations differently in different cultural or economic circumstances.[102] This inherent flexibility is likely to be all the more true for platforms, which have further room for adaptation created by the essentially unprecedented process of translating state-based free expression norms into the fresh and different context of the online environment.

Worse still, such variance will depend on the particular platforms' product, technical capacity, and other contextual factors. As Benesch observes, "If [IHRL] come[s] to guide online content moderation, different platforms may derive quite different rules from them."[103] Indeed, as the non-governmental organization Article 19 has observed:

> A degree of variation is inherent to [IHRL]. . . . In the case of [platforms], this margin of appreciation in the application of international standards will allow the application of standards in a specific country context (as is generally the case with international standards), and the differentiation between different companies and their respective products (e.g., Facebook is different from Twitter), including the liberty of a company to adopt stricter restrictions on freedom of expression to accommodate their own editorial choices (although it should be clear that market dominance would result in a narrower margin of appreciation in this respect).[104]

Even disregarding regional instruments, different rights protected by IHRL will often come into conflict. As the Human Rights Review of Facebook's Oversight Board observed, in cases where rights conflict "it will be important for the Oversight Board to have a clear approach to 'counterbalancing' different human rights," and recommending a form of structured proportionality testing for doing so.[105] A civil rights audit admonished Facebook, for example, for taking an unduly "selective view of free expression as Facebook's most cherished value," without accounting for impacts on other rights.[106] Ultimately, there will always be room for balancing values and making choices. These choices will rest with the front-line decision-makers in content moderation: platform policy makers.

---

102. ANDREW LEGG, THE MARGIN OF APPRECIATION IN INTERNATIONAL HUMAN RIGHTS LAW: DEFERENCE AND PROPORTIONALITY 6 (Vaughan Lowe, Dan Sarooshi & Stefan Talmon eds., 2012).

103. BENESCH, *supra* note 84, at 17.

104. ARTICLE 19, SOCIAL MEDIA COUNCILS: CONSULTATION PAPER 13 (2019).

105. BUS. FOR SOC. RESP., HUMAN RIGHTS REVIEW: FACEBOOK OVERSIGHT BOARD 35 (2019).

106. LAURA MURPHY, FACEBOOK'S CIVIL RIGHTS AUDIT - FINAL REPORT 9 (2020).

This highlights a distinctive aspect of IHRL's role in this context as opposed to that which is seemingly directly contemplated in the UNGPs: the role of a platform "respecting" IHRL in the context of a platform is adjudicatory. A platform recognizes and determines disputes about the rights of its users, as balanced against the rights of other users or society's interests as a whole (and with its own commercial interests always lurking in the background). This is quite a different task from those most obviously addressed by the UNGPs (such as physical business operations, supply chain issues, or labour rights), which do not discuss what businesses should do when human rights obligations are more nebulous or involve balancing various rights and interests against each other.[107]

In short, platform lawyers will always have a lot of leeway when they translate IHRL into the context of their particular product and any particular case.

It is somewhat abstract to discuss this indeterminacy in generic terms, and so I will return to this point below in the context of the specific examples of hate speech and election interference. But the overarching point is true for many, if not most, of the decisions platforms make, especially the most difficult and controversial ones: there are multiple possible rights-respecting answers within the framework offered by IHRL.

D. Co-optation

The indeterminacy of IHRL creates room for its co-optation by platforms, rather than their being constrained by it. This is similar to charges of "bluewashing" that have attended initiatives such as the United Nations Global Compact, where given the lack of monitoring and enforcement, members are able to enjoy goodwill and legitimacy benefits without costly changes.[108] In the context of information technologies specifically, a similar controversy arose in the early days of the Global Network Initiative (GNI), which was formed in 2008 in the wake of scandals in which both Yahoo! and Google were found to be involved in human rights violations in China. Years later, however, the Edward Snowden revelations showed that GNI verification mechanisms had failed to reveal the extent of data collection and sharing by many GNI members with the U.S. government.[109] The Electronic Frontier Foundation (EFF), a founding member of the GNI, resigned as a result.[110]

Platforms do make commitments to uphold human rights and have been hiring personnel to rectify their long-standing lack of human rights expertise.[111] But

107. *Id.* at 10. Indeed, the only reference to such "balancing" is the "difficult balancing decisions" states have to make to "reconcile different societal needs."

108. *See, e.g.*, Berliner & Prakash, *supra* note 8, at 115, 118.

109. LAIDLAW, *supra* note 13, at 104–08.

110. Letter from Danny O'Brien & Jillian C. York, Elec. Frontier Found., to Glob. Network Initiative 1 (Oct. 9, 2013), https://www.eff.org/document/gni-resignation-letter (explaining decision to withdraw from GNI).

111. MACKINNON, *supra* note 32, at 138–39.

some reports suggest that such initiatives are window-dressing, with those involved marginalized within the companies or not given sufficient resources and authority to make a real difference.[112]

Facebook has taken the commendable step of commissioning human rights impact assessments (HRIAs) in developing countries.[113] But even these are telling. The four reports released so far combined do not match the length or depth of the Civil Rights Audit Facebook commissioned and released about its impact in the United States.[114] The recommendations were unsurprising and did not necessarily require an in-depth understanding of IHRL; they mostly amounted to simple due diligence before entering a volatile market. Facebook's responses to the HRIAs essentially promised to devote more resources to these markets, hire more staff, provide greater language support, and pointed to a number of other initiatives that it already had in train (such as its Oversight Board and limiting forwarding in WhatsApp).[115] This series of measures can be described as common sense. That is not to say that they are not welcome, just that the role of IHRL could be described as marginal or a redundancy over seriously living up to the motto "don't be evil." The reports received minimal press coverage in the West (or, perhaps, anywhere), and there were no timelines given for taking next steps, nor has there been any follow-up.

By casting the reports as HRIAs, Facebook has co-opted the language of IHRL without any substantial changes in operations or IHRL-specific commitments beyond common sense due diligence. In the end, "such documents are only worth what the receiver makes of them."[116]

These same "bluewashing" risks could attend the adoption of IHRL standards in content moderation more broadly. As Sander argues, "[g]iven [the] complexity [of applying IHRL to platforms], 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."[117] So far, so true. Even as Facebook has said it has grounded its community standards in IHRL, researchers observed that those writing the rules

112. Nitasha Tiku, *A Top Google Exec Pushed the Company to Commit to Human Rights. Then Google Pushed Him Oout, He Says.*, WASH. POST (Jan. 2, 2020), https://www.washingtonpost.com/technology/2020/01/02/top-google-exec-pushed-company-commit-human-rights-then-google-pushed-him-out-he-says/.

113. Miranda Sissons & Alex Warofka, *An Update on Facebook's Human Rights Work in Asia and Around the World*, FACEBOOK NEWSROOM (May 12, 2020), https://about.fb.com/news/2020/05/human-rights-work-in-asia/.

114. *See* MURPHY, *supra* note 106 (there was also a substantial interim report).

115. Sissons & Warofka, *supra* note 113, at 1, 4–5.

116. LAIDLAW, *supra* note 13, at 168 (noting similar concerns about a human rights audit of the Internet Watch Foundation).

117. Sander, *supra* note 2, at 1006.

"were not observed to refer to concrete human rights norms."[118] The gap between rhetoric and practice remains.

But the risk of co-optation is even greater than in other contexts because it runs both ways. Not only could companies co-opt IHRL rhetoric to their own advantage, but such co-optation could bleed back into IHRL itself and influence its development. As Aswad observes, "[i]f companies begin applying Article 19(3) in their content moderation operations and take up the Special Rapporteur's call to produce 'case law,' there could be an active fountain of new 'jurisprudence' involving the ICCPR's speech protections, which could influence the direction of international freedom of expression rights."[119] The development of IHRL through state practice, authoritative interpretation by treaty bodies, and relevant experts can be slow and ad hoc. By contrast, content moderation systems work at an extremely high pace and volume. The gap between these agile "move fast" systems and the creaky IHRL structures is vast. As such, the extent to which private companies may shape "international normative developments and discourse on freedom of expression" is unclear,[120] but could be profound.

E. Lack of Competency

Even setting aside concerns about bad faith co-optation of IHRL by platforms and assuming best intentions, platforms do not have the legitimacy or competence to determine IHRL obligations alone.

Platforms have no democratic legitimacy to draw upon when conducting the fraught task of determining whether a particular restriction on freedom of expression is necessary and proportionate.[121] But this weakness applies to all content moderation. There are more specific competence problems created in the application of IHRL.

First, it is unclear which interests platforms should take into account as justifying limitations on expression. Article 19 provides that freedom of expression should only be limited "to protect rights or reputations of others; national security; public order; public health or morals."[122] But as Kaye noted in his report on online hate speech, "companies are not in the position of governments to assess threats to national security and public order, and hate speech restrictions on those grounds should be based not on company assessment but legal orders from a State."[123] If this implies that companies *are* in a position to evaluate restrictions on speech for

118. KETTEMANN & SCHULZ, *supra* note 38, at 32.
119. Aswad, *supra* note 75, at 64.
120. Parmar, *supra* note 35, at 6.
121. Kaye, *supra* note 1.
122. ICCPR, *supra* note 27, art. 19(3), at 178.
123. Kaye, *supra* note 30, ¶ 47(b)15.

other purposes such as rights or reputations of others and public health or morals,[124] experience so far is not promising.

Platforms do not necessarily have all the information necessary to evaluate the extent to which harm is done to someone's reputation off-platform, nor do they have any particular competence or legitimacy to determine matters of public health or morals. The Covid-19 pandemic has been a stark illustration of how even ostensibly science-based public health determinations are fraught, often subject to conflicting guidance and politicization.[125] As to public morals, platforms are at best limited judges of such social matters, and this is all the more true in the case of quintessentially Silicon Valley-based companies purporting to determine the morals of users around the world.[126]

Second, compounding matters, it is unclear the extent to which platforms can take into account their own legitimate business and free speech interests in deciding what to allow on their services. Unlike First Amendment doctrine, IHRL speech protections do not extend to corporations.[127] But it is not clear that denying corporate interests entirely or making them completely subservient to individual speech rights would be the best outcome, legally or normatively.

Legally, as Molly Land notes, while human rights institutions have tended to simply "equate public and private censorship," it "cannot be the case that every content moderation decision made by every digital platform should be subject to human rights scrutiny."[128] While the extent may be unclear, it should be relatively uncontroversial to assert that clearly private actors can limit, curate, and privilege content and viewpoints in a way that a state should not. Moderation of clearly legal content is an important part of what platforms are: without it, platforms would quickly find themselves overrun with spam, adult content, and other merely 'unpleasant' content that would diminish the value of their products to users and,

124. These are the other legitimate grounds for limitation of a right under article 19 of the ICCPR.

125. There are many examples, but perhaps the starkest have been the politicization around mask-wearing and hydroxychloroquine, both exacerbated by unclear or flat-out mistaken messaging from the authoritative sources that platforms said they would look to in determining what content to restrict.

126. *See, e.g.*, Olivia Solon, *Inside Facebook's Efforts to Stop Revenge Porn Before it Spreads*, NBC NEWS, (Nov. 19, 2019), https://www.nbcnews.com/tech/social-media/inside-facebook-s-efforts-stop-revenge-porn-it-spreads-n1083631 ("Davis gave the example of a woman in India who reported a photo in which she was fully clothed in a pool with a fully clothed man. 'Within her culture and family that would not be acceptable behavior,' Davis said. 'The photo was shared intentionally with her family and employer to harass her.'"); Chris Marsden, Trisha Meyer & Ian Brown, *Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?*, 36 COMPUT. L. & SEC. (2020) ("Executives in California, ex-politicians such as Nick Clegg, or thousands of badly-paid contractors hired off the internet, from the Philippines or India, cannot regulate European fake news: it has to be Europeans.").

127. Aswad, *supra* note 75, at 40.

128. Land, *supra* note 57, at 292.

in some cases, make it unusable.[129] But when a platform does so, "what values should guide its decision?"[130] Business interests cannot be an entirely illegitimate consideration—indeed, companies have a fiduciary duty to maximize stockholder value.[131]

Furthermore, there is good reason to think that preserving a diversity of platform approaches is overall more beneficial than compelled conformity. Some platforms may choose to try to "inspire creativity and bring joy,"[132] while others will seek to "give everyone the power to create and share ideas and information instantly without barriers."[133] A degree of experimentation and choice should be welcomed.[134] State-based bodies of law do not have a good answer for how to account for this. Molly Land and Emily Laidlaw have argued, for example, that obligations should be calibrated to the extent of a platform's dominance or influence on democracy.[135] IHRL already supports the notion that a stronger showing of the necessity of restrictions is required when there is a lack of viable alternative channels for communication.[136] This is echoed in the quote from Article 19, above.[137]

But even this leaves things fairly indeterminate and fails to answer crucial questions such as the extent to which intermediaries can account for any unique business mission or user preference in removing content, for example, offensive content.[138] Laidlaw concluded that IHRL just doesn't "quite fit" with what platforms are doing.[139] For many questions about the obligations of gatekeepers, "there is little guidance in [IHRL]"[140] and the UNGPs "raise just as many questions as they answer."[141] As such, it remains true that, as Kaye observed in 2016, it is 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."[142]

---

129. GILLESPIE, *supra* note 13, at 5.

130. Land, *supra* note 57, at 292.

131. Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries*, 133 HARV. L. REV. 497, 503–04 (2019).

132. *About TikTok*, https://www.tiktok.com/about?lang=en (last visited Sept. 25, 2020).

133. *Twitter, Inc. - Contact - FAQ*, TWITTER, https://investor.twitterinc.com/contact/faq/default.aspx (last visited Sept. 25, 2020).

134. Sander, *supra* note 2, at 981.

135. Land, *supra* note 57, at 304.

136. Sander, *supra* note 2, at 981 n. 184.

137. ICCPR, *supra* note 27, art. 19.

138. Molly K. Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, 60 VA. J. INT'L L. 3631, 3932 (2020).

139. LAIDLAW, *supra* note 13, at 111.

140. *Id.* at 89.

141. *Id.* at 96.

142. David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 55, U.N. Doc. A/HRC/32/38 (May 11, 2016).

In short, in restricting expression, platforms "might base these choices on balancing public interests or their own private interests, but in neither case are they well-placed to be trusted arbiters,"[143] and IHRL does not offer concrete guidance.

F. Unearned Legitimacy Dividends

One of the strongest arguments in favor of IHRL in content moderation is the value not of its substantive norms, which the preceding sections have argued remain contested and indeterminate in many cases, but its procedural requirements.[144] As outlined above, on this argument, the constraining and legitimating force of IHRL comes not from it forcing platforms to arrive at the right *decisions*, but by guiding them to arrive at decisions in the right *way*. The requirements of legality, legitimacy, necessity, and proportionality are intrinsically valuable, independent from the ultimate substantive outcomes.

This aligns with arguments I have made elsewhere about the importance of transparent and public reasoning in platform rule and decision-making in legitimating their exercise of substantial power over speech.[145] But here's the rub: there is nothing particular to IHRL about these requirements. The structured nature of the tests and what they require overlaps to a very large extent with structured proportionality testing that exists in many constitutional systems and is the globally dominant form of judicial review.[146] The requirements of a clear, precise, and transparent statement of a rule that is justified in the pursuance of a legitimate purpose are generic rule of law and due process requirements that can be found in diverse areas of law.[147] The tripartite test in Article 19(3) is a useful framework for facilitating this process, but there is nothing uniquely powerful about this framework as distinct from other forms of structured proportionality testing.

---

143. Douek, *supra* note 10, at 461.

144. *See supra* Part II(E).

145. *See* Douek, *supra* note 31; Douek, *supra* note 62; Douek, *supra* note 10.

146. *See, e.g.*, Alec Stone Sweet & Jud Mathews, *Proportionality Balancing and Global Constitutionalism*, 47 COLUM. J. TRANSNAT'L. L. 72 (2008); Vicki C. Jackson, *Constitutional Law in an Age of Proportionality*, 124 YALE L.J. 3094 (2015); PROPORTIONALITY: NEW FRONTIERS, NEW CHALLENGES (Vicki C. Jackson & Mark V. Tushnet eds., 2017); Bernhard Schlink, *Proportionality in Constitutional Law: Why Everywhere But Here?*, 22 DUKE J. COMP. & INT'L L. 291 (2012); AHARON BARAK, PROPORTIONALITY: CONSTITUTIONAL RIGHTS AND THEIR LIMITATIONS (Doron Kalir trans., 2012); ALEC STONE SWEET & JUD MATHEWS, PROPORTIONALITY BALANCING AND CONSTITUTIONAL GOVERNANCE: A COMPARATIVE AND GLOBAL APPROACH (2019); Jamal Greene, *Foreword: Rights as Trumps?*, 132 HARV. L. REV. 28 (2018); Grégoire C. N. Webber, *Proportionality, Balancing, and the Cult of Constitutional Rights Scholarship*, 23 CAN. J. L. & JURIS. 179 (2010); Evelyn Douek, *All Out of Proportion: The Ongoing Disagreement About Structured Proportionality in Australia*, 47 FED. L. REV. 551 (2019).

147. For an argument that many of these principles underlie administrative law, see Cass R. Sunstein & Adrian Vermeule, *The Morality of Administrative Law*, 131 HARV. L. REV. 1924 (2018).

The structured nature of the test, setting out distinct steps, is important and provides a disciplining and rationalizing effect absent in open-ended balancing.[148] But it is both the strength and weakness of this kind of analysis that it can—and must—be adapted to the particular context in which it is being used. Such structures remain "framework[s] that must be filled with content."[149] This is a strength because, as the preceding sections have demonstrated, there are significant adaptations that need to be made to the unique context of social media platforms. But it is also a weakness because this very adaptability is the thrust of one of the major critiques of proportionality testing in global constitutionalism: it is too indeterminate.[150]

My own view is that the structure and transparency of this kind of reasoning are intrinsically valuable, regardless of the fact that it is adaptable and does not constrain the decision-maker to particular outcomes or even particular considerations.[151] But it remains important to acknowledge this indeterminacy openly for exactly the same reason that the transparent process of reasoning is important in the first place: to make "the calculus behind an opinion explicit so that it can be seen and criticized."[152] We should be cautious of letting platforms clothe themselves in the language of IHRL and accrue legitimacy dividends merely for meeting bare minimum transparency and justification requirements.

G. New Paradigms of Rights

Finally, and crucially (as I have argued elsewhere), technology has created the capacity for both more speech and more speech governance than any time in history and this requires a paradigm shift in thinking about online speech rights.[153] This paradigm requires a systemic, rather than individualistic, view of rights.

IHRL, like most constitutional systems, takes an individualistic view of speech rights. Article 19 of the ICCPR begins "*[e]veryone* shall have the right to freedom of expression."[154] Cases focus on particular individuals and even particular utterances. But the scale of the major social media platforms makes such thinking ill-fitted to modern online speech governance.

---

148. Dieter Grimm, *Proportionality in Canadian and German Constitutional Jurisprudence*, 57 U. TORONTO L.J. 383, 397 (2007); Frederick Schauer, *Balancing, Subsumption, and the Constraining Role of Legal Text*, *in* INSTITUTIONALIZED REASON: THE JURISPRUDENCE OF ROBERT ALEXY 307, 308–09 (Matthias Klatt ed., 2012).

149. BARAK, *supra* note 146, at 489–90.

150. Grimm, *supra* note 148, at 397.

151. *See* Douek, *supra* note 10.

152. STEPHEN BREYER, THE COURT AND THE WORLD: AMERICAN LAW AND THE NEW GLOBAL REALITIES 257 (2015).

153. This section draws on Douek, *supra* note 10.

154. ICCPR, *supra* note 27, art. 19 (emphasis added).

Errors in content moderation at scale are inevitable,[155] and therefore the more pertinent question in the design of a content moderation system is what kinds of errors to err on the side of.[156] IHRL does not lend itself to this kind of analysis. Instead, it suggests that rights-respecting online speech governance "requires more than just optimizing a speech-regulation system for a small quantity of error; it requires individualized determinations by independent arbiters."[157]

But this is infeasible at internet scale, which makes literally billions of pieces of content potentially governable every day. Kaye anticipated this argument, noting that "[s]ome may argue that it will be time-consuming and costly to allow appeals on every content action. But companies could work with one another and civil society to explore scalable solutions such as company-specific or industry-wide ombudsman programmes."[158]

In my view, such scalable solutions not only practically require a more systemic view of rights as opposed to an individualistic one, but also provide a lens that is normatively preferable. A system that acknowledges the inevitability of error at the outset will often lead to fewer errors in the long run than trying to design a system on the assumption that all errors can be caught and corrected.[159] Procedural requirements are a case in point. Determining what process is due in any case, as discussed above, is a quintessentially contextual judgment.[160] Fundamentally,

> the very nature of the due process inquiry indicates that the fundamental fairness of a particular procedure does not turn on the result obtained in any individual case; rather, "procedural due process rules are shaped by the risk of error inherent in the truth-finding process as applied to the generality of cases."[161]

My point here is not to resolve the question but simply to note that IHRL alone does not. Content moderation system design requires unprecedented thinking about speech governance and system design which IHRL can provide some guardrails for, but not yet ultimately answer. It needs new tools to do so, and we should not put social media companies alone in charge of crafting and applying those tools.

This applies more broadly to the "proportionality" assessment as applied to content moderation. One benefit of the online environment is that platforms can

---

155. *See* Evelyn Douek, *COVID-19 and Social Media Content Moderation*, LAWFARE (Mar. 25, 2020, 1:10 PM), https://www.lawfareblog.com/covid-19-and-social-media-content-moderation; Mike Masnick, *Masnick's Impossibility Theorem: Content Moderation at Scale is Impossible to do Well*, TECHDIRT (Nov. 20, 2019, 9:31 AM), https://www.techdirt.com/articles/20191111/23032743367/masnicks-impossibility-theorem-content-moderation-scale-is-impossible-to-do-well.shtml.

156. Douek, *supra* note 10.

157. Llansó, *supra* note 61, at 4 (citing Nunziato, *supra* note 60).

158. Kaye, *supra* note 1, at 18.

159. FREDERICK F. SCHAUER, PROFILES, PROBABILITIES, AND STEREOTYPES 23 (2009).

160. Mathews v. Eldridge*,* 424 U.S. 319 (1976).

161. Walters v. Nat'l Ass'n of Radiation Survivors, 473 U.S. 305, 321 (1985) (quoting Mathews, 424 U.S. 319, 344 (1976)).

develop far more nuanced remedies than have traditionally been available to governments.[162] Far beyond the false binary of choosing to leave content up or take it down, they can choose to label it, amplify it, suppress it, demonetize it, or engage in counter-messaging.[163] IHRL encourages, if not mandates, that platforms explore these options by requiring any measure restricting expression be necessary, in the sense that it is the least intrusive instrument.[164] Again, to date it does not offer any guidance beyond this general command, placing essentially all content moderation decisions into a grey zone.[165] A jurisprudence can develop over time, but the platform lawyer charged with making these determinations today is left on their own.

III. HARD CASES

The thrust of my critique has been that using IHRL in content moderation sounds good in theory but has limits in practice. This part illustrates this argument by reference to some concrete examples.

I have already noted that IHRL does not give much guidance on a number of categories of content moderation that may be mundane and not attract much attention but actually make up the bulk of platform decision-making, such as spam, nudity, obscenity, bullying, and self-harm. Most of this content would be protected under IHRL as applied to states, but if platforms were forced to carry it, it would dramatically change the nature of their services and in some cases make them practically unusable. For example, Facebook removed 35.7 million pieces of adult nudity in the second quarter of 2020,[166] and this is even given a strongly established and well-known norm against posting such content. Again, my point is not that IHRL cannot adjust to these requirements, and no doubt it will do so as leading thinkers attack the task. I only mean to underline that at present, what IHRL requires in such a case is an important open question, likely turning on a number of contextual factors which vary platform-by-platform. It, therefore, is limited as a source of constraint.

Nevertheless, these are not typically the questions that dominate discussions around content moderation, human rights, and international law. Here, I examine two high-profile and controversial examples, hate speech and foreign election interference, and show that IHRL would not necessarily offer determinate answers that would help lead to greater perceived legitimacy for platform decisions.

---

162. Kaye, *supra* note 30, at 16–17.

163. Douek, *supra* note 10, at 26–27.

164. Hum. Rts. Comm., *General Comment No. 34, Article 19: Freedoms of Opinion and Expression*, ¶¶ 3, 34, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011).

165. Sander, *supra* note 2, at 988 ("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.").

166. FACEBOOK, COMMUNITY STANDARDS ENFORCEMENT REPORT 5 (2020)*, Q2 2020*, FACEBOOK TRANSPARENCY, https://transparency.facebook.com/community-standards-enforcement#adult-nudity-and-sexual-activity.

A. Hate Speech & Holocaust Denial

In a way, all content moderation debates—perhaps even all free speech debates—are debates about hate speech, such is the shadow the topic casts over the field. As one of the hardest categories of content to define in the abstract and apply in practice, concerns about how to draw the line between permissible and impermissible speech and "who decides," loom especially large. This subpart, therefore, asks whether looking to IHRL makes these questions easier.

As the NGO Article 19 has noted, "'[h]ate speech' is an emotive concept which has no universally accepted definition in [IHRL]."[167] Indeed, the words "hate speech" do not appear at all in IHRL treaties.[168] As a result, "[i]nternational doctrine and practice relating to prohibition of hate speech remain uneven."[169] It is also "where the chasm between US and international approaches is the greatest,"[170] potentially making adoption of IHRL in this area especially controversial.

The Rabat Plan of Action provides a useful and more detailed guide, to be sure, proposing a six-part threshold test for what should constitute unlawful incitement under Article 20.[171] But some of these elements, such as intent and imminence of harm, will be hard to determine and always require in-the-moment judgments for which few prior guardrails can be provided. This is the perennial problem with crafting rules for hate speech, which necessarily require highly contextual judgments.[172] In short, in most cases a platform will need to defend its application of generalized IHRL to the specific context on its own terms.

Take an example of a case where IHRL might have, on its face, resolved some persistent legitimacy problems for platforms. Facebook in particular had long come under public pressure about its decision not to remove content denying the historical fact of the Holocaust.[173] Removal of Holocaust denial was one of the key demands of the high-profile #StopHateForProfit campaign in 2020.[174] Facebook has stood firm for years in the face of public outrage. Aswad argued that in this

167. ARTICLE 19, SELF-REGULATION AND 'HATE SPEECH' ON SOCIAL MEDIA PLATFORMS 6 (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.

168. BENESCH, *supra* note 84, at 16.

169. Cleveland, *supra* note 78, at 225.

170. *Id.* at 210.

171. U.N. High Comm'r for Hum. Rts., *Report on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred*, ¶ 1, U.N. Doc. A/HRC/22/17/ADD. 4, (Jan. 11, 2013).

172. Richard Ashby Wilson & Molly K. Land, *Hate Speech on Social Media: Towards a Context-Specific Content Moderation Policy*, 52 CONN. L. REV. 1, 47 (2020).

173. Ezra Klein, *The Controversy Over Mark Zuckerberg's Comments on Holocaust Denial, Explained*, VOX (July 20, 2018, 11:30 AM), https://www.vox.com/explainers/2018/7/20/17590694/mark-zuckerberg-facebook-holocaust-denial-recode.

174. STOP HATE FOR PROFIT, https://www.stophateforprofit.org/ (last visited Oct. 11, 2020).

context pointing to IHRL could have given Facebook the rationalization for its decision that it struggled to find.[175]

But even this apparently straightforward application of IHRL is not obviously the right answer. First, IHRL is not as clear as it could be on even this relatively confined question. The Human Rights Committee famously found in 1996 that a French conviction of an academic Holocaust denier did not violate Article 19 of the ICCPR,[176] although in 2011 the Committee appeared to confine the decision to its facts in saying that "laws that penalize the expression of opinions about historical facts are incompatible with . . . respect for freedom of opinion and expression."[177] Platforms would need to justify their reconciliation of these two conflicting authorities. This may not be difficult: it could be explained as an evolution of IHRL doctrine rather than an inconsistency.

But even assuming IHRL speaks clearly on the issue, do the rationales for preventing a state from deciding on an official view of history apply equally to a social media platform that is not seeking to criminalize such speech or prevent academic research on the topic? What about the other different characteristics of online speech? Kaye himself has raised the question, "[g]iven the speed, reach, and capacity for disinformation online, should our definition of 'online hate speech' be different from offline hate speech?"[178] In evaluating Holocaust denial, should platforms acknowledge Europe's unique historical experience?

Ultimately, the position under IHRL was not enough for Facebook to resist public pressure or otherwise prevent it from reaching a different conclusion with respect to Holocaust denial. In late 2020, Facebook[179] and Twitter[180] finally joined YouTube[181] in banning such content. Do these policies breach the platforms' obligations under IHRL? If so, what about Reddit—does its status as a smaller platform give it greater license to "quarantine" Holocaust denial?[182] Is there a different balancing exercise for search engines, or even search functions within all social media platforms, based on a right to truth and memory as a collective right,

175. Aswad, *supra* note 75, at 66.; *see* Hum. Rts. Comm., *supra* note 164, ¶ ¶ 35, 49.

176. Faurisson v. France, CCPR/C/58/D/550/1993, ¶ 10 (Dec. 16, 1996).

177. Hum. Rts. Comm., *supra* note 164, ¶ 49.

178. David Kaye, *Four Questions about Online Hate Speech*, MEDIUM (Aug. 12, 2019), https://medium.com/@dkisaway/four-questions-about-online-hate-speech-ae3e0a134472.

179. Monika Bickert, *Removing Holocaust Denial Content*, FACEBOOK NEWSROOM (Oct. 12, 2020), https://about.fb.com/news/2020/10/removing-holocaust-denial-content/.

180. Jacob Kastrenakes, *Twitter Will Ban Holocaust Denial Posts, Following Facebook*, VERGE (Oct. 14, 2020), https://www.theverge.com/2020/10/14/21516468/twitter-holocaust-denial-banned-facebook-policy.

181. Paresh Dave, *YouTube Reversal Bans Holocaust Hoaxers*, *Stops Pay for Borderline Creators*, REUTERS (June 5, 2019, 9:51 AM), https://www.reuters.com/article/us-alphabet-youtube-hatespeech-idUSKCN1T623X.

182. *ADL Statement on Reddit's New Quarantine Policy for Holocaust Denial*, ANTI-DEFAMATION LEAGUE (Sept. 28, 2018), https://www.adl.org/news/press-releases/adl-statement-on-reddits-new-quarantine-policy-for-holocaust-denial.

as Deirdre Mulligan and Daniel S. Griffin argue, even as they acknowledge that the exact form of this right and its implementation is thorny and complicated?[183]

IHRL also does not directly answer how a platform should evaluate which local laws it should comply with as legitimate deviations from ICCPR and which it should not. For example, while the European Court of Human Rights found in 2019 that the conviction of a German politician for Holocaust denial did not violate the European Convention,[184] Germany cannot rely on ECHR jurisprudence to justify its deviation from the ICCPR standard that forbids such bans.[185] On the other hand, the UNGPs require platforms to comply with local law.[186] So when is a deviation from ICCPR standards merely a legitimate local law with which companies should comply, and when is it an instance where IHRL should give them 'stiffened spines' to resist authoritarian demands? There might be easy cases, and perhaps Germany's laws on Holocaust denial fall into this bucket given its unique historical experience, but there will be many more hard ones.

To me, it is not obvious that IHRL offers a clear and uncontestable answer to these questions: a platform will still need to defend its decision on its own terms, in its own context, and based on the resources and technical capacity it has to enforce any such rules and the resulting error rates. Kaye himself seems to envision exactly this kind of diversity, stating "[w]here company rules vary from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation."[187]

Things get worse when we move from the more confined category of denial of historical facts to hate speech more broadly. All the same issues apply but are compounded by the fact that the underlying IHRL norms, and especially the relationship between Articles 19 and 20, are not always clear. This is further compounded in the online environment where culture can evolve rapidly, be subject to differing interpretations in different communities, and coded language is especially prominent. Perhaps no example illustrates this as well as Pepe the Frog, which over the course of its brief life has been a comic book character, an anti-Semitic hate symbol, and a pro-Democracy message in Hong Kong.[188] The hypothetical platform Head of Policy may get some basic guidance from IHRL on

183. Deirdre K. Mulligan & Daniel S. Griffin, *Rescripting Search to Respect the Right to Truth*, 2 GEO. L. TECH. REV. 557, 577–78, 582 (2018).

184. Pastörs v. Germany, App. No. 55225/14 (Oct. 3, 2019), http://hudoc.echr.coe.int/eng?i=001-196148.

185. Aswad, *supra* note 75, at 58 n. 152.

186. Aswad, *supra* note 77, at n. 59.

187. Kaye, *supra* note 30, at 15–16.

188. Brittan Heller, Opinion, *Is This Frog a Hate Symbol or Not?*, N.Y. TIMES (Dec. 24, 2019), https://www.nytimes.com/2019/12/24/opinion/pepe-frog-hate-speech.html.

how to treat such situations but cannot truly be said to be constrained by it in deciding what to do.[189]

In the context of rapidly evolving language and norms, the question of error-choice is especially acute. Does a platform have sufficient technical capacity and resources to apply a distinction between hateful slurs and attempts by target communities to reclaim the language, for example? If not, should the platform err on the side of over-enforcing or under-enforcing its policy against hate speech, acknowledging that while the former would perhaps be a precautionary approach to protect marginalized communities it is also at odds with traditional thinking about freedom of expression that generally errs on the side of giving free speech "breathing space."[190] IHRL does not speak in these terms of error choice, but rather in the language of individual cases. Practically, however, these considerations pervade platform decision-making.[191]

There is, of course, a far richer and growing body of IHRL jurisprudence and scholarship on the topic of hate speech than I could do justice to here, not in small part because it is such a vexing issue.[192] I do not mean to diminish the importance of platforms engaging with and paying heed to this work. It can play a meaningful role in informing their decision-making process and drawing platform lawyers' attention to relevant considerations. My point is simply that when the rubber hits the road, this work cannot answer specific questions or help those lawyers choose between various alternative courses of action. The constraining value, and therefore the legitimating force we should endow it with, will be limited.

B. Election Interference

Foreign election interference on social media is the scandal that perhaps more than any other kicked off the "techlash" and plausibly played a significant role in creating the public pressure that led to platforms embracing IHRL as an attempted solution to their legitimacy deficits. And yet, as Jens David Ohlin has commented, "international lawyers seem to be at a loss for how to understand the particular harm posed by [Russian] interference."[193] The ICCPR of course specifies that everyone has the right to freedom of expression "*regardless of frontiers*."[194] This applies even to falsehoods—IHRL does not include an exception enabling restrictions on freedom

189. *See* BENESCH, *supra* note 84, at 16–17.

190. New York Times Co. v. Sullivan, 376 U.S. 254, 271–72 (1964).

191. Douek, *supra* note 10.

192. *See, e.g.*, Kaye, *supra* note 30; Hum. Rts. Comm., *supra* note 164; Wilson & Land, *supra* note 172; Arun, *supra* note 93; Cleveland, *supra* note 78; BENESCH, *supra* note 84; ARTICLE 19, 'HATE SPEECH' EXPLAINED: A TOOLKIT (2015), https://www.article19.org/data/files/medialibrary/38231/'Hate-Speech'-Explained---A-Toolkit-%282015-Edition%29.pdf; as well as many of the pieces cited throughout this article.

193. Jens David Ohlin, *Election Interference: The Real Harm and The Only Solution* (Cornell Legal Stud. Rsch. Paper No. 18-50, 2018), https://ssrn.com/abstract=3276940.

194. ICCPR, *supra* note 27, art. 19 (emphasis added).

of expression merely because content is "false."[195] Indeed, in the years following WWII, the United States and other democratic states "monotonously" refused Soviet Union demands for a treaty outlawing international war propaganda on the grounds that this would jeopardize freedom of speech.[196] The best cure for international propaganda, these states maintained, was more, not less, freedom of information.[197] Thus, cross-border speech, even by states, is protected by IHRL. The extent to which even state-sponsored doxing and disinformation operations are in violation of IHRL is complicated by contestation around the scope of IHRL's extraterritorial application.[198] Outlawing speech on the basis of its foreignness alone would be a violation of international law; on the other hand, disclosure obligations, so that actors cannot conceal or mislead users as to their identity, are in the best counter-speech tradition.[199] But here, IHRL has no work to do: every major platform is committed—in theory—to the detection and removal of foreign election interference.

The harder question is what level of coordination and platform manipulation, whether foreign or domestic, is impermissible interference with an individual's right to seek and receive information and ideas.[200] But on this question, IHRL offers little guidance: again, no legal system does yet. This is a new frontier. The internet and social media have blurred traditional lines between "coordinated efforts" to game a system and the "genuine" output of users: "[m]ost contributions to the web are somewhere in the middle, where people in some way coordinate their efforts in order to help make their content visible to a search engine, out of a 'genuine' desire for it to be seen."[201] Indeed, "defining what is coordination and what is inauthentic is far from a value-free judgment call. . . . Coordination and authenticity are not binary states but matters of degree, and this ambiguity will be exploited by actors of all stripes."[202]

---

195. KAYE, SPEECH POLICE, *supra* note 11, at 94.

196. JOHN B. WHITTON & ARTHUR LARSON, PROPAGANDA: TOWARDS DISARMAMENT IN THE WAR OF WORDS 234 (1964).

197. *Id.* at 241.

198. Barrie Sander, *Democracy Under The Influence: Paradigms of State Responsibility for Cyber Influence Operations on Elections*, 18 CHINESE J. INT'L L. 1, 39 (2019).

199. Ohlin, *supra* note 193.

200. *See, e.g.*, U.N. Special Rapporteur on Freedom of Opinion and Expression et. al., *Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda*, U.N. Doc. FOM.GAL/3/17 (Mar. 3, 2017), https://www.osce.org/files/f/documents/6/8/302796.pdf ("[D]isinformation and propaganda are often designed and implemented so as to mislead a population, as well as to interfere with the public's right to know and the right of individuals to seek and receive, as well as to impart, information and ideas of all kinds, regardless of frontiers, protected under international legal guarantees of the rights to freedom of expression and to hold opinions.").

201. Tarleton Gillespie, *Algorithmically Recognizable: Santorum's Google Problem, and Google's Santorum Problem*, 20 INFO. COMM. & SOC'Y 63, 67 (2017).

202. Evelyn Douek, *What Do Platforms Think "Coordinated Inauthentic Behavior" Actually Means?*, SLATE (July 2, 2020, 5:26 PM), https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html.

Kaye suggested in 2017 that "[i]t was not entirely clear if international norms spoke to how governments should address" online propaganda,[203] and there has been no progress in answering that question since then. Indeed, as more and more state actors engage in precisely this kind of activity both internationally and domestically, being able to distill any agreed norms as a matter of IHRL seems doubtful at best.

IV. ADVANCING IHRL IN ONLINE SPEECH GOVERNANCE

Having argued both that IHRL is the least-worst option for content moderation baselines *and* that it is inadequate, it is somewhat incumbent on me to offer a path forward. A full account is beyond my ambit here,[204] but I believe there are four takeaways from my argument so far.

First, how IHRL applies to platform governance and online speech is still highly uncertain and developing. There should be nothing surprising about this: no legal system has good answers yet for how to deal with the fundamental changes that the internet has wrought to societies' speech ecosystems. International law is no exception. IHRL is still young and general, but "just as the vaguely worded First Amendment has crystallized into more concrete rules, so too can international law."[205] Constructing systems of free expression takes time, experimentation, and incremental development. The messiness of content moderation and the role of IHRL within it is not a failing but simply a feature of the current stage of evolution of both.

Second, however, is that facilitating the growth of both from this period of adolescence to maturity requires being candid about the areas in which such development is necessary. Too much of the discourse around IHRL in content moderation smooths over the complexities or weaknesses that will necessarily be inherent in this process of adaptation of IHRL norms to a fundamentally new context, which leaves them vulnerable to exploitation by the very actors that such norms are supposed to constrain (namely, platforms).

Third, for the movement to imbue content moderation with IHRL to be meaningful, it cannot again just outsource the interpretation and application of these norms to the private companies themselves in the same way that content moderation has been outsourced so far—this will only allow for the agenda's co-optation. IHRL can provide an important source of standards for content moderation but only if, as Benesch argues, "properly interpreted and explained by experts."[206] I agree, but I would add a further caveat to Benesch's caveat: human rights expertise will not be enough to shore up the legitimacy of any such

203. KAYE, SPEECH POLICE, *supra* note 11, at 93.
204. How convenient!
205. Douek, *supra* note 1.
206. BENESCH, *supra* note 84, at 15.

interpretations. There still needs to be a push to get multi-stakeholder buy-in in order for any interpretation to carry normative weight. Therefore, the movement needs to provide an institutional framework to check and balance platforms' development of IHRL and place platforms in conversation with other stakeholders.

As discussed above,[207] one of the key benefits of IHRL in this context is that it can provide a common vocabulary for content moderation debates so that even as rules are contested and "the participants in these debates plainly disagree about which policies promote the public good[,] . . . there is value to putting them in conversation with one another."[208] That is, even if, as the proceeding sections have suggested, the *substantive* constraints on platforms under IHRL would be minimal, international law as an argumentative practice has independent value.[209] But the important caveat is that for argumentative practice to be successful, participants must actually be *in conversation*. Creating legitimacy and accountability through argumentative practice requires an institutional structure that facilitates exactly this kind of argument and contestation. The use of IHRL by platforms, unmoored from any institutional hierarchy or constraints and not embedded in a broader conversation, cannot facilitate this process.

There are some proposals for independent or quasi-independent institutions to check platforms' application of and compliance with IHRL, such as social media councils or Facebook's Oversight Board.[210] These proposals are already in train and most embrace something in this vein.[211] These initiatives hold promise,[212] but cannot solve all the concerns raised here. Facebook's Oversight Board is the most developed, but as an institution of Facebook's creation cannot fully cure the lack of legitimacy and competency in applying IHRL, despite the presence of some IHRL experts amongst its members. This is especially so if the intention is for platforms to adopt readings of IHRL that contradict regional bodies' interpretations of IHRL: the authority of the Board to demand that Facebook ignore such state-backed institutions is highly questionable. The multi-stakeholder model of a social media council including government representatives with democratic mandates may hold

207. *See supra* Part II(C).

208. Hakimi, *supra* note 46, at 1304.

209. *Id.* at 1301.

210. ARTICLE 19, *supra* note 104; Douek, *supra* note 31.

211. *See, e.g.*, Sander, *supra* note 2, at 988, 1002–03 (describing the importance of an institutional supporting framework); GLOB. PARTNERS DIGIT., A RIGHTS-RESPECTING MODEL OF ONLINE CONTENT REGULATION BY PLATFORMS 38, 26 (2018), https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf (proposing an independent Online Platform Standards Oversight Body); TRANSATLANTIC HIGH LEVEL WORKING GROUP ON CONTENT MODERATION ONLINE AND FREEDOM OF EXPRESSION, FREEDOM AND ACCOUNTABILITY: A TRANSATLANTIC FRAMEWORK FOR MODERATING SPEECH ONLINE 44, 26–28 (2020), https://cdn.annenbergpublicpolicycenter.org/wp-content/uploads/2020/07/Freedom_and_Accountability_TWG_Final_Report.pdf (embracing the need for social media councils and e-courts).

212. As to the promise and potential pitfalls of the Oversight Board, see Evelyn Douek, *"What Kind of Oversight Board Have You Given Us?,"* U. CHI. L. REV. ONLINE (2020), https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek/.

more potential in this respect, but its form is still inchoate and its legitimacy in demanding interpretations that diverge from state-based bodies' rulings still unclear.

Again, none of this is to diminish the value or importance of these discursive benefits provided by stakeholders engaged in the project of adapting IHRL to the work of platforms. Such argumentation and deliberation are inherently valuable as society adapts to this new information age and muddles through the new challenges it poses for us. But on the more specific question with which this Article is concerned—the extent to which IHRL can decide and provide legitimation for platform policies and decisions in hard cases, especially to those that are prone to doubt platform bona fides or IHRL in general—it is important to acknowledge that the offering is more limited.

Accordingly, the argument of this Article has been that there needs to be some established mechanism for checking and contesting platforms' use of IHRL that incorporates state consent and therefore has democratic legitimacy. In the absence of such a mechanism, platforms' embrace of IHRL cannot be declared a meaningful victory. Otherwise, there is a risk of creating a situation where platforms' use of IHRL is praised when it aligns with outcomes its advocates like and dismissed when it differs. This itself could, in the long term, undermine the larger project of creating an IHRL-based framework for guiding content moderation by suggesting that participants only favor it when they agree with the outcomes ostensibly issued in its name.

This possibility has darker potential costs. One of the most important uses of IHRL for platforms is providing normative backing and legitimation for companies to push back against government overreach[213] in an effort to weather the storm of rising digital authoritarianism. Allowing companies to deploy IHRL language in cases where it does little substantive work and is indeterminate, providing little more than vocabulary for the action companies would take in any event, could weaken its normative force for when it is really needed: as a bulwark against state oppression.

CONCLUSION

Social media platforms should respect and uphold the rights of their users, regardless of their status as private profit-maximizing businesses. Their rules and decisions have profound impacts on individuals and societies, and they should exercise that power in a public-regarding way and be held accountable for doing so. This Article should not be read as a call to throw the baby out with the bathwater. The significant normative force of IHRL can be seen in Facebook's reliance on it in challenging a Thai government order to block a private group because it was

213. *See supra* Part II(D).

critical of the monarchy,[214] or the decision of several major platforms to suspend government requests for user data in Hong Kong after a new national security law infringed on freedom of expression and other human rights in the city.[215] There are unfortunately many such examples and no doubt there will be more in the future. But this context of pushing back against state overreach is very different from a platform purporting to create new IHRL directly in its own content moderation decisions.

The proposal for this latter use of IHRL in content moderation has garnered remarkable momentum and support in an extremely short period of time. This is a credit to the many strengths of the idea and the important advocacy of those who have pushed for it. But even as companies start to acknowledge these calls and say they are adopting advocates' agenda, we should not be lulled into a false sense of security. Acknowledging the limitations and complexities of IHRL in content moderation will ultimately serve to strengthen the long-term endeavor of holding companies to account for their impact on human rights. If platforms want the legitimacy dividends associated with respecting IHRL, they should pay for them. We must demand such payment and create the institutions necessary to ensure that IHRL in content moderation serves the interests of users and society rather than being co-opted by platforms to their own ends.

214. Pavin Chachavalpongpun, Opinion, *An Entire Generation in Thailand is Counting on Facebook to Do the Right Thing*, WASH. POST (Aug. 28, 2020, 6:11 AM), https://www.washingtonpost.com/opinions/2020/08/28/an-entire-generation-thailand-is-counting-facebook-do-right-thing/.

215. Paul Mozur, *TikTok to Withdraw From Hong Kong as Tech Giants Halt Data Requests*, N.Y. TIMES (July 6, 2020), https://www.nytimes.com/2020/07/06/technology/tiktok-google-facebook-twitter-hong-kong.html.