VOLUME 64, NUMBER 1, WINTER 2023

Expert Governance of Online Speech

Brenda Dvoskin*

In a world of fundamental disagreements about how social media companies should govern speech, it is striking that nearly everyone agrees that online speech governance should be based on human rights. The human rights project for content moderation proposes that social media platforms align their own internal speech policies with international human rights law. It seeks, I argue, a system of expert governance: one in which a corporate technocracy applies a set of exogenous principles imagined as objective and global. Ultimately, this governance model shifts power to experts under the illusion of empowering the people.

To support these claims, this Article unveils the intellectual work that scholars, U.N. bodies, and the Facebook Oversight Board are doing to portray international human rights law as an objective synthesis of the global public interest. The Article analyzes how they have recreated several dimensions of international law. A salient example is their new reading of the U.N. Guiding Principles on Business and Human Rights. According to a recent interpretation, companies are expected to align their content policies with international law. But this interpretation widely diverges from the text and the original meaning of the instrument. The Article also examines other tools the project uses such as creating boundaries between local facts and normative work and framing normative questions as technical challenges. Overall, the Article provides a deep dive into the toolkit that scholars, advocates, and the Facebook Oversight Board have developed to date to pursue a system of expert governance of online speech.

* Postdoctoral fellow, Georgetown University Law Center; Doctoral candidate, Harvard Law School; Affiliate, Berkman Klein Center for Internet & Society. I am deeply thankful for insightful comments from Dunstan Allison-Hope, Chinamyi Arun, Nicole Bassoff, Yochai Benkler, Elettra Bietti, Alejandro Chehtman, Evelyn Douek, Noah Feldman, Sheila Jasanoff, Gerald Neuman, Peter Stern, Thomas Streinz, Gali Racabi, Malcolm Rogge, and Yiran Zhang. I am also grateful to the participants at the Critical Exploration of Human Rights Conference at the University College Dublin Center for Human Rights, the Ideas Lunch at the Information Society Project at Yale Law School, the 2022 Global Meeting on Law and Society, and the Science, Technology and Society Circle at the Harvard Kennedy School. Thank you to Gabriella Papper, Deniz Aktaş, Sophia Poole, and Sara Raza for their work editing this piece. The Article reflects developments and decisions through August 2022.

Introduction

In April 2021, the Facebook Oversight Board (hereinafter "Board") concluded that general prohibitions on content depicting people in blackface are incompatible with international human rights law (hereinafter "IHRL").[1] Simultaneously, however, the Board decided that Facebook had "met its human rights responsibilities" by adopting precisely that prohibition.[2] The case originated with a video featuring two adults representing Zwarte Piet (or "Black Pete"), the companion of Saint Nicholas in the Dutch folklore. Zwarte Piet has been widely reported as racist because people have traditionally portrayed it by putting on blackface.[3] In August 2020, Facebook prohibited "caricatures of Black people in the form of blackface."[4] Facebook deleted the post for violating that rule.

This decision ought to be somewhat perplexing. How did Facebook meet its human rights responsibilities by adopting a rule incompatible with human rights law? How did the Board explain a conclusion that, at least at first sight, was palpably incoherent? The Board articulated various argu-

1. *Case decision 2021-002-FB-UA*, Oversight Bd. (Apr. 13, 2021), https://oversightboard.com/decision/FB-S6NRTDAJ/ [https://perma.cc/5PDM-VD6F] ("The Board notes international human rights law would not allow a state to impose a general prohibition on blackface through criminal or civil sanctions except under the conditions foreseen in ICCPR Article 20, para. 2 and Article 19, para. 3 . . . Expression that does not reach this threshold may still raise concern in terms of tolerance, civility and respect for others, but would not be necessary or proportionate for a state to restrict . . . ."). In all cases, I refer to the English version of the Board's decisions.

2. *Id.*

3. U.N. Comm. on the Elimination of Racial Discrimination, *Concluding observations on the nineteenth to twenty-first periodic reports of the Netherlands*, ¶¶ 15-17, U.N. Doc. CERD/C/NLD/CO/19-21 (Aug. 28, 2015).

4. Guy Rosen, *Community Standards Enforcement Report, August 2020*, Meta (Aug. 11, 2020), https://about.fb.com/news/2020/08/community-standards-enforcement-report-aug-2020/ [https://perma.cc/53BN-NGFN].

ments. One strand of the reasoning noted that international law protects deeply offensive speech, suggesting that depictions of Zwarte Piet might be protected. However, the Board determined that the international human rights question turned on whether featuring blackface caused "objective harm" or "subjective offense." The Board enumerated the many experts who had verified that content featuring blackface is discriminatory, reinforces harmful stereotypes, can impact individuals' self-esteem, and can contribute to an environment of intimidation.[5] Experts' reports were considered evidence of "*objective* harm."[6]

This Article aims to explain why the Board as well as many scholars and advocates find that IHRL is a promising framework to make decisions regarding the governance of online speech. It dissects the argumentative mazes they must navigate to justify both that nonstate actors should implement international law and that the rules they propose to moderate online content are *objective* applications of that framework.

Over the last decade, scholars have often compared social media companies to states due to the control they exercise over the public sphere.[7] It is therefore unsurprising that many proposals to deal with corporate power take inspiration from state institutions.[8] In that vein, what I call the IHRL project for content moderation proposes that social media platforms align their own internal rules to govern speech with IHRL. In other words, the IHRL project posits that companies should adopt the international legal framework to protect individuals' rights to freedom of expression from state interference in order to protect individuals from corporate power.[9]

Intuitively, IHRL is an attractive framework for corporate actors to govern speech on social media. It is global like social media platforms and it offers well-respected standards to guide content moderation in these quasi-

5. *Case decision 2021-002-FB-UA*, *supra* note 1.

6. *Id.* (emphasis added).

7. *See generally infra* Section II.A.2.

8. *See* Lex Gill et al., *Towards Digital Constitutionalism? Mapping Attempts to Craft an Internet Bill of Rights*, 80 Int'l Commc'n Gazette 302 (2018) (discussing proposals to craft a constitution for the internet); Kate Klonick & Thomas Kadri, *Facebook v. Sullivan: Public Figures and Newsworthiness in Online Speech*, 83 S. Cal. L. Rev. 37, 38 (2019) (arguing that platforms act as legislature, executive, and judiciary, without any separation of powers); Noah Feldman, *Facebook Supreme Court: A Governance Solution*, *in* Global Feedback and Input on the Facebook Oversight Board for Content Decisions Appendix, https://about.fb.com/wp-content/uploads/2019/06/oversight-board-consultation-report-appendix.pdf [https://perma.cc/W37W-R5H5] (discussing the usefulness of judicial models in the context of social media); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. Rev. 27, 71-78 (2019) (drawing from administrative law to reorganize the structure of social media companies); The Santa Clara Principles, https://santaclaraprinciples.org/ [https://perma.cc/4MPL-STUC] (extending due process rights to users vis-à-vis platforms).

9. 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*, U.N. Doc. A/HCR/38/35 (Apr. 6, 2018) (hereinafter "SR Report 2018").

public spaces.[10] The IHRL project proposes to address the democratic deficits of corporate online speech governance not by suggesting a more participatory institutional structure, but by offering a set of global substantive rules that purport to reflect the public interest, the common good, or globally shared values.[11] Indeed, blurring the lines between states and corporations could lead to seeing users as political subjects with an interest in participating in the making of policy. Instead, the IHRL project proposes that experts implement IHRL as an already available formulation of social interests.

The project has gained more attention and traction since David Kaye proposed it in 2018 during his tenure as U.N. Special Rapporteur.[12] He recommended that large social media platforms adopt IHRL as their own default rules to moderate content.[13] Also in 2018, Mark Zuckerberg announced that Facebook would create an overseeing body composed of independent experts to review some of Facebook's content decisions and to make recommendations on how the company could improve its content governance.[14] From 2021 to the present, the Board has applied the IHRL framework to its decisions. The Board has been the main executor of the IHRL project to date.[15]

IHRL does not assure a panacea, but it makes appealing promises. Its main purpose is to put the public interest front and center of online speech governance. Kaye asserts: "[i]t's time to put individual and democratic rights at the center of corporate content moderation."[16] The project challenges the idea that no global set of rules to regulate speech exists. In his 2018 report, Kaye stated, "[t]he founder of Facebook recently expressed his hope for a process in which the company 'could more accurately reflect the values of the community in different places.' That process, and the relevant standards, can be found in human rights law."[17] Thus, the IHRL project

10. *See* Sejal Parmar, *Facebook's Oversight Board: A Meaningful Turn Toward 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/ [https://perma.cc/LTW4-E6HS].

11. David Kaye, Speech Police: The Global Struggle to Govern the Internet 18 (2019).

12. SR Report 2018, *supra* note 9. *See* Rebecca MacKinnon, Consent of the Networked: The Worldwide Struggle for Internet Freedom (2012) (pioneering the importance of human rights in the context of social media).

13. SR Report 2018, *supra* note 9. *See* Evelyn Douek, *U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for 'Human Rights by Default'*, Lawfare (June 6, 2018), https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default [https://perma.cc/X3HT-5CDA].

14. Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, Facebook (Nov. 15, 2018), https://m.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634/ [https://perma.cc/KD2H-7CWU].

15. *Announcing the Oversight Board's first case decisions*, Oversight Bd. (Jan. 2021), https://oversightboard.com/news/165523235084273-announcing-the-oversight-board-s-first-case-decisions/ [https://perma.cc/JRG9-BBZ7].

16. Kaye, *supra* note 11, at 18.

17. SR Report 2018, *supra* note 9, ¶ 41 (internal footnotes omitted).

claims that it can provide a basis to moderate content that already reflects the global public interest.[18]

By positing that we already know what rules reflect the global public interest, the IHRL project vests legitimacy on new "new governors": human rights experts.[19] The Board's members or companies' in-house human rights directors—those who know these universal rules—can implement these rules on social media. The experts' democratic credentials would proceed not from democratic politics but from the fact that they deduce their decisions from IHRL.

I call this project a system of expert governance. Its legitimacy comes from assuring the public that IHRL is an objective account of the public interest or global values, capable of producing rules and decisions that also reflect those values. Objectivity in this context means the opposite of individual policy preferences. It is also the opposite of politics. I do not imply that human rights themselves are not a political commitment. The IHRL project's claim to objectivity stems from portraying that commitment as universal and pre-existing to the decisionmaking moment. Thus, the system is objective because it is supposed not to favor anyone's specific standpoint.

As Sheila Jasanoff, a pioneer in the field of Science and Technology Studies, says, objectivity takes hard work.[20] Performing as objective governors (judges often play this role) requires construing principles as being already available and agreed on, as well as framing decisions as reasonable derivations from them. In the case about Black Pete, the Board had to present IHRL as the framework it was compelled to rely on and to claim that it deduced its decision from that framework, even though the Board acknowledged that the exact opposite conclusion was also possible.

The main purpose of this Article is to make that work visible. It offers an analysis of the tools that scholars, advocates, and the Board have developed to build that claim to objectivity. By examining how objectivity is performed, this Article makes four contributions.

First, it unveils the intellectual work that has gone into presenting IHRL as a global set of rules appropriate for governing speech online. Sometimes the effort toward objectivity is explicit. Like in the Black Pete decision, experts might explain the work they are doing to construe the principles that allegedly bind them and the parameters for diverging from them. Most often, the work is done in the dark.[21] This should be unsurprising to legal

18. *Id.*, ¶ 42.

19. *See generally* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 Harv. L. Rev. 1598 (2018).

20. Sheila Jasanoff, *The Practices of Objectivity in Regulatory Science*, *in* Social Knowledge in the Making 307, 308 (Charles Camic, Neil Gross & Micèle Lamont eds., 2011).

21. *See* McDonald v. City of Chicago, 561 U.S. 742, 805 (2010) (Scalia, J., concurring) ("Justice Stevens abhors a system in which 'majorities or powerful interest groups always get their way,'. . .but replaces it with a system in which unelected and life-tenured judges always get their way. That such usurpation is effected unabashedly. . . —with 'the judge's cards . . . laid on the table,'. . .—makes it even

scholars. Just as judicial courts have developed theories of interpretation that legitimize their work,[22] proponents of the IHRL project have done a great deal of intellectual work to build the principles that ought to constrain experts' decisionmaking process. For example, the thought leaders of the project have argued that the United Nations Guiding Principles for Business and Human Rights (hereinafter "UNGPs") set the expectation that companies will align their policies with IHRL.[23] This interpretation serves the purpose of presenting this instrument as an exogenous basis for the IHRL project. Looking at how creative this interpretation is and how widely it diverges from previous interpretations illuminates the work done behind principles that are portrayed as objective.[24]

Second, by bringing this intellectual work into the light, the Article extends long-standing critiques of courts to the Board.[25] The Board presents IHRL as a reflection of a global agreement or set of values about how to regulate speech. That conceptualization of IHRL allows the Board to hide away its power, that is, to perform as an enforcer of the global public interest instead of as a political actor making policy. The risk, however, is that this version of the IHRL project undermines its own purpose. While the goal at first was to bridge the divide between public and market actors and include the public in the governance of social media, the outcome is to reaffirm that the participation of the public is ultimately not necessary because its interests are already embedded in IHRL.

Third, looking at how corporate actors implement international law illustrates the consequences of importing state institutions in the hope of addressing the problematic aspects of corporate power. When incorporated into nonstate actors, state institutions go through radical transformation to fit the existing structure and purposes of corporations. This often leads to

---

worse. In a vibrant democracy, *usurpation should have to be accomplished in the dark*." (internal citations omitted; emphasis added). I am grateful to Libby Adler for highlighting this quote.

22. *See* Noah Feldman, *Written Statement to the Presidential Commission of the Supreme Court of the United States Hearing on "The Contemporary Debate over Supreme Court Reform: Origins and Perspectives"*, THE WHITE HOUSE (June 30, 2021), https://www.whitehouse.gov/wp-content/uploads/2021/06/Feldman-Presidential-Commission-6-25-21.pdf [https://perma.cc/4V6A-RSXU].

23. *See* Evelyn Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 34 (2018); David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Mandate of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 1, 2 U.N. Doc. OL OTH 24/2019 (May 1, 2019); *Facebook Oversight Board: Recommendations for human rights-focused oversight*, ARTICLE 19 (Mar. 27, 2019), https://www.article19.org/resources/facebook-oversight-board-recommendations-for-human-rights-focused-oversight/ [https://perma.cc/399N-JUVV] ("While Facebook is not legally bound by international human rights laws, the UN Guiding Principles on Business and Human Rights set out responsibilities that companies like Facebook have to respect human rights. This means ensuring that their Terms of Service and Community Standards are fully in line with human rights laws").

24. *See infra* Section II.A.1.

25. *See* Jeremy Waldron, *The Core of the Case against Judicial Review*, 115 YALE L.J. 1346 (2006); Mark Tushnet, *Following the Rules Laid down: A Critique of Interpretivism and Neutral Principles*, 96 HARV. L. REV. 781 (1983).

puzzling outcomes.[26] This Article examines what gets lost in translation when corporate actors become international lawmakers.

Finally, the Article reviews many of the Board's decisions to date. Because the Board has become the main executor of the IHRL project, its decisions are a useful site to see the project (or at least one version of it) at work. Thus, the Article reviews many of the Board's first cases, although this review is by no means a comprehensive review of all of the Board's work.

I proceed as follows: the first section offers a description of the IHRL project and how it advances a system of expert governance. The second section examines the rationales offered to justify that companies should adopt IHRL as their corporate speech codes. These rationales show, I argue, that the IHRL project is rooted in an ideal of objectivity and expertise. The third section looks at the work that goes into maintaining the claim to objectivity when adjudicating speech within the IHRL framework. It examines several tools experts use to legitimize their own power. The fourth section digs into the oscillation between shifting power to experts and creating new avenues for public participation. It looks at how these two mechanisms of governance compete within the IHRL project. The last section reflects on the risks of the IHRL project as executed to date and imagines more promising futures.

I. The International Human Rights Law Project for Online Speech Governance

A large consensus among companies, scholars, and advocates has emerged around the idea that regulation coming from either states or private actors should be based on human rights. Twitter's guiding principles for government regulation enunciate that the internet should be built on "the protection of human rights."[27] Facebook's principles for online content regulation state that "the most important elements of any system will be due regard to each of the human rights and values at stake."[28] Companies say that human rights should inform not only government regulation but also their own internal rules.[29] Likewise, many scholars agree that content moderation

26. *See* Brenda Dvoskin, *Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, 67 Vill. L. Rev. 447 (2022) (exploring the distributional consequences of corporate adoption of administrative law principles for public participation).

27. *Protecting the Open Internet: Regulatory Principles for Policy Makers*, Twitter, https://cdn.cms-twdigitalassets.com/content/dam/about-twitter/en/our-priorities/open-internet.pdf [https://perma.cc/5QJW-27QG].

28. Monika Bickert, *Online Content Regulation: Charting a Way Forward*, Facebook, https://about.fb.com/wp-content/uploads/2020/02/Charting-A-Way-Forward_Online-Content-Regulation-White-Paper-1.pdf [https://perma.cc/NHB7-9GJL].

29. *See, e.g., Defending and Respecting the Rights of People Using Our Service*, Twitter, https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice [https://perma.cc/2RYJ-T44D]; Jack Dorsey (@jack), Twitter (Aug. 10, 2018, 9:58 AM), https://twitter.com/jack/status/1027962500438843397 [https://perma.cc/VZA3-7E43]; Miranda Sissons, *Our Commitment to Human*

should follow a human rights approach.[30] Multiple civil society organizations focusing on online speech identify themselves as human rights organizations.[31] An overwhelming number of reports, proposals, and commentary from civil society and advocates argue that it would be desirable to base content moderation on human rights.[32]

The prevalence of the human rights language may only indicate a broad consensus that content moderation must respond to the public interest. In some cases, human rights functions as an undefined term to refer to the

---

Rights, META (Mar. 16, 2021), https://about.fb.com/news/2021/03/our-commitment-to-human-rights/ [https://perma.cc/S9AX-NNKY]; *Facebook Community Standards*, META, https://transparency.fb.com/policies/community-standards/ [https://perma.cc/A4LC-NXQS] ("These standards are based on feedback from people and the advice of experts in fields like technology, public safety and human rights."); Monika Bickert, *Updating the Values That Inform Our Community Standards*, META (Sept. 12, 2019), https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/ [https://perma.cc/5UK6-QN38].

30. *See, e.g.,* NICOLAS SUZOR, LAWLESS: THE SECRET RULES THAT GOVERN OUR DIGITAL LIVES 125 (2019); HUMAN RIGHTS IN THE AGE OF PLATFORMS (Rikke Frank Jørgensen ed., 2019); Dinah Pokempner, *Regulating Online Speech: Keeping Humans, and Human Rights, at the Core*, *in* FREE SPEECH IN THE DIGITAL AGE 224 (Susan Brison & Katharine Gelber eds., 2019); 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 (2020); Kate Jones, *Online Disinformation and Political Discourse: Applying a Human Rights Framework*, CHATHAM HOUSE (Nov. 6, 2019), https://www.chathamhouse.org/2019/11/online-disinformation-and-political-discourse-applying-human-rights-framework [https://perma.cc/RRB9-XEA3]; Susan Benesch, PROPOSALS FOR IMPROVED REGULATION OF HARMFUL ONLINE CONTENT, 1, 10, https://dangerousspeech.org/wp-content/uploads/2020/07/Proposals-for-Improved-Regulation-of-Harmful-Online-Content-Formatted-v5.2.2.pdf [https://perma.cc/DP8T-5YJJ]. *But see,* Brenda Dvoskin, *International Human Rights Law Is Not Enough to Fix Content Moderation's Legitimacy Crisis*, MEDIUM (Sept. 16, 2020), https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd [https://perma.cc/53VZ-S8H9]; Evelyn Douek, *The Limits of International Law in Content Moderation*, 6 U.C. IRVINE J. INT'L, TRANSNAT'L & COMPAR. L. 37 (2021); Rachel Griffin, Rethinking Rights in Social Media Governance: Human Rights, Ideology and Inequality (Dec. 15, 2021) (unpublished manuscript) (on file with author).

31. *See, e.g.,* GLOB. NETWORK INITIATIVE, https://globalnetworkinitiative.org/ [https://perma.cc/2T2Q-MGJQ]; HUM. RTS. WATCH, https://www.hrw.org/ [https://perma.cc/YTR9-KKH4]; AMNESTY INT'L, https://www.amnesty.org/ [https://perma.cc/EN47-R946]; RANKING DIGITAL RTS., https://rankingdigitalrights.org/ [https://perma.cc/94G8-CRQW]; ARTICLE 19, https://www.article19.org/ [https://perma.cc/A43R-7L39].

32. *See, e.g., Corporate Speech Controls,* ELECTRONIC FRONTIER FOUNDATION, https://www.eff.org/issues/corporate-speech-controls [https://perma.cc/DGW8-M7RM] ("we believe that they should promote free expression and transparency and base moderation practices and policies on human rights norms."); Jillian C. York & Corynne McSherry, *Content Moderation is Broken. Let Us Count the Ways.*, ELECTRONIC FRONTIER FOUNDATION (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-count-ways [https://perma.cc/97VJ-ZUJE]; Eliška Pírková & Javier Pallero, *Twenty-Six Recommendations on Content Governance: A Guide for Lawmakers, Regulators, and Company Policy Makers*, ACCESS NOW 1, 35 (2020), https://www.accessnow.org/cms/assets/uploads/2020/03/Recommendations-On-Content-Governance-digital.pdf [https://perma.cc/A777-4U7T]; Parmar, *supra* note 10; Laura Murphy, *Facebook's Civil Rights Audit–Final Report*, FACEBOOK 1, 9 (July 8, 2020), https://about.fb.com/wp-content/uploads/2020/07/Civil-Rights-Audit-Final-Report.pdf [https://perma.cc/K9U4-BGK4]; Charles Bradley & Richard Wingfield, *A Rights-Respecting Model of Online Content Regulation by Platforms*, GLOBAL PARTNERS DIGITAL (May 2018), https://www.gp-digital.org/wp-content/uploads/2018/05/A-rights-respecting-model-of-online-content-regulation-by-platforms.pdf [https://perma.cc/5E8F-VFYE]; Emma Llansó, *CDT's Comments to Facebook Oversight Board on 2021-001-FB-FBR (Case Regarding Suspension of Trump's Account)*, CTR. FOR DEMOCRACY & TECH. (Feb. 11, 2021), https://cdt.org/insights/cdts-comments-to-facebook-oversight-board-on-2021-001-fb-fbr-case-regarding-suspension-of-trumps-account/ [https://perma.cc/W8JQ-RVWX].

interests, preferences or agreements of society as opposed to the commercial interests of private entities.[33] Other times, companies and advocates refer explicitly to IHRL, a body of treaties and authoritative interpretations with established—although often vague or contested—norms and conditions for states to regulate speech.

IHRL comprises several treaties, instruments, and authoritative interpretations, but Article 19 of the International Covenant on Civil and Political Rights (hereinafter "ICCPR") is the epicenter of the IHRL project for content moderation. This provision sets out a tripartite test to evaluate restrictions on freedom of expression. The test from Article 19 includes three requirements: rules must be prescribed by law (legality), must have a legitimate aim (legitimacy), and must be necessary for that aim (necessity). Succinctly, the first requirement means that restrictions on speech must be provided by law, that is, they must be enacted within a country's domestic legal system.[34] To meet this requirement, restrictions must also be clear and precise enough to give appropriate notice to the public of what speech is not allowed.[35] Second, a restriction on speech must have one of the public interest aims enumerated in Article 19(3): the protection of the rights or reputations of others, national security, public order, public health, or morals.[36] Finally, for speech restrictions to be considered necessary, they must be the least intrusive means to achieve their legitimate aim and they must be proportionate to the interest they are designed to protect.[37]

The IHRL project for content moderation promises important and valuable contributions.[38] Scholars emphasize IHRL's procedural mandates.[39] They argue that, in accordance with IHRL, companies have to be transparent about how they moderate content and have to offer their users robust appeal processes to exercise their due process rights.[40] IHRL can also function as a loose normative agreement among multiple constituencies that can serve as a

33. Rikke Frank Jørgensen, *What Platforms Mean When They Talk About Human Rights*, 9 Pol'y & Internet 280 (2017) (showing how companies present themselves as public spaces). *See also* Amy Kapczynski, *The Right to Medicines in an Age of Neoliberalism*, 10 Human. J. 79, 85 (2019) (describing the polycentric uses of the term "human rights.").

34. U.N. Hum. Rts. Comm., *General Comment No. 34, Article 19: freedoms of opinion and expression*, ¶ 24, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011) (hereinafter "General Comment 34"); David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Rep. of the Special Rapporteur on the Promotion and Prot. of the Right to Freedom of Op. and Expression*, ¶ 12, U.N. Doc. A/71/373 (Sept. 6, 2016) (hereinafter "SR Report 2016").

35. General Comment 34, *supra* note 34, ¶ 25.

36. International Covenant on Civil and Political Rights, art. 19, 999 U.N.T.S. 171, 178 (entered into force Mar. 23, 1976) (hereinafter "ICCPR").

37. General Comment 34, *supra* note 34, ¶ 34.

38. *See* Douek, *supra* note 30, at 41-50 (providing an overview of how international law can improve online speech governance).

39. *See, e.g.,* Sander, *supra* note 30, at 988.

40. *See, e.g.,* 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).

starting point for designing content moderation.[41] Similarly, taking inspiration from an existing legal system can be a useful tool for corporations to make decisions about online speech. For that purpose, there is no better alternative than IHRL.[42]

The IHRL project is an ongoing development carried out by multiple scholars and advocacy groups. Therefore, different versions of the project compete.[43] The main ambivalence within the project is exactly what guidance companies are expected to follow.

On the one hand, companies are called to follow IHRL, which comprises the U.N. treaties, regional treaties, international custom, and the authoritative interpretations by courts, special rapporteurs, and other international bodies. In that sense, scholars have focused their attention on translating the substantive rules of international law to make them usable by corporate actors.[44]

Within this version of the project, scholars disagree on whether the project should focus exclusively on translating U.N. level treaties or should also include regional human rights treaties and national interpretations of international instruments.[45] David Kaye argues that regional treaties and local interpretations of international instruments can provide important guidance,[46] although other authors believe it is wiser to limit the scope of the project to U.N. treaties.[47] At the same time, while in some cases the goal has been to operationalize IHRL into granular rules that can effectively guide speech governance, scholars have also called for multi-stakeholder conversations to determine what these rules should be.[48]

On the other hand, a variation of the IHRL project suggests that content moderation policies should meet the requirements of legality, legitimacy, and proportionality of Article 19(3) of the ICCPR to assess the reasonableness of their rules without necessarily adhering to existing determinations of

41. Jacob Mchangama et al., *A Framework of First Reference: Decoding a Human Rights Approach to Content Moderation in the Era of "Platformization",* THE FUTURE OF FREE SPEECH (2021), https://futurefreespeech.com/wp-content/uploads/2021/11/Report_A-framework-of-first-reference.pdf [https://perma.cc/BVP8-MJNQ].

42. Douek, *supra* note 30, at 49.

43. *See Social Media Councils: Consultation Paper*, ARTICLE 19, (2019), https://www.article19.org/wp-content/uploads/2019/06/A19-SMC-Consultation-paper-2019-v05.pdf [https://perma.cc/P6WD-28ZM] (surveying competing proposals to align content moderation with international human rights law).

44. *See, e.g.,* Aswad, *supra* note 23; Sander, *supra* note 30, at 969 ("Arguably the biggest challenge, however, *resides in the translation of general human rights principles into particular rules, processes and procedures tailored to the platform moderation context*.") (emphasis in original); *id*. at 971 ("*{T}ranslation* from the State to the corporate context of platform moderation is likely to pose a number of challenges in practice.") (emphasis added).

45. *See infra* Section III.B.

46. David Kaye, *A New Constitution for Content Moderation,* MEDIUM (June 25, 2019), https://onezero.medium.com/a-new-constitution-for-content-moderation-6249af611bdf [https://perma.cc/2FSR-VZ9U].

47. *See, e.g.,* Aswad, *supra* note 23, at 44; *infra* Section III.B. (discussing the role of regional and U.N. instruments in the IHRL project).

48. *See* Aswad, *supra* note 23, 35.

which governmental restrictions meet these conditions. Accordingly, the expectation is that companies should be able to justify their moderation policies as reasonable, but international law would not impose significant restrictions on which policies are acceptable.[49]

This ambivalence between aligning content moderation with IHRL and reducing IHRL to a generic proportionality test (and all possible intermediate possibilities of asking companies to adhere to some but not all international law standards) makes the IHRL project flexible and resilient. For example, if the objection is that IHRL is too vague to offer guidance for content moderation, the response is that companies can look to the vast authoritative interpretations advanced by international and local tribunals, treaty bodies, and others. But if the objection is that those sources are incoherent or not appropriate for private content moderation, the response is that companies do not need to comply with international law.

This Article traces this ambivalence and accounts for how different trends are negotiated within the project. I pay particular attention, however, to how the Board is carrying out the IHRL project because it is the most relevant actor implementing it. Currently, the Board has 23 members with diverse geographical and professional backgrounds.[50] Many, but not all of these members, have strong expertise in IHRL. In response to the project's ambiguity, the Board's decisions offer a concrete description of what content moderation can look like when it is based on human rights standards. Therefore, it could be that some of the tools, implementation challenges, and concerns that I describe here are specific to this version of the project.

Despite the variations, the IHRL project for corporate content moderation in all of its versions, I believe, envisions an ideal of expert and objective governance. Performing as objective decisionmakers is a strategy commonly pursued by judges and bodies of experts.[51] The field of science and technology studies describes various mechanisms that experts, including judges, use to exercise power while maintaining the appearance of objectivity and democratic buy-in.[52] One such mechanism is to construct exogenous and neutral governing standards. Because these standards purport not to reflect anyone's

49. Douek, *supra* note 30, at 64 (expressing concern that platforms might "clothe themselves in the language of IHRL and accrue legitimacy dividends merely for meeting bare minimum transparency and justification requirements.").

50. *Meet the Board*, Oversight Bd., https://www.oversightboard.com/meet-the-board/ [https://perma.cc/EE8T-AKY6]; Evelyn Douek, *What Kind of Board Have You Given Us?*, Chicago L. Rev. Blog (May 11, 2020), https://lawreviewblog.uchicago.edu/2020/05/11/fb-oversight-board-edouek/ [https://perma.cc/Y5NF-87DS] (analyzing the structure of the Board, its founding documents, and its initial members).

51. *See generally* Aziza Ahmed, *Medical Evidence and Expertise in Abortion Jurisprudence*, 41 Am. J.L. & Med. 85 (2015) (exploring how medical evidence and expertise are portrayed as neutral and objective in the context of abortion access); Michael Klarman, From Jim Crow to Civil Rights: The Supreme Court and the Struggle for Racial Equality (2005) (analyzing how judges reconstruct the law to reconcile their understanding of what is legally required with their moral views).

52. Sheila Jasanoff, *Subjects of Reason: Good, Markets and Competing Imaginaries of Global Governance*, 4 London Rev. Int'l. L. 361, 363 (2016).

standpoint, Sheila Jasanoff describes them as representing a view from nowhere.[53] An alternative mechanism is to build principles that are presented as neutral because they represent a view from everywhere.[54] These principles are imagined as the result of a participatory process to reach consensus and agreement, so the outcome is attributed to all the participants.[55] The IHRL project relies on both kinds of objectivity. IHRL is both portrayed as self-evidently good and as reflecting a universal consensus.[56]

The hypothesis of the project is that IHRL can function as an already available account of the interests or will of global society. However, as I hope will become apparent in the next sections, the idea of IHRL as a synthesis of the global public interest is an illusion.[57] While the vagueness of IHRL shows that IHRL cannot define those interests with any precision, attempts to construct more precise rules are equally misguided. The public's will is fragmented and in permanent conflict.[58] Still, these efforts imagine that what is needed is more time, thinking, and conversations to find an intelligible, coherent, and precise enough synthesis of the public's will. In turn, the project assumes it is possible to hold experts accountable to the public by asking them to follow that synthesis.

The danger of believing that it is possible to find an account of the will of the global society is that it enables experts to assert that they are acting on behalf of everyone (or no one). If experts were to recognize that their decisions benefit only the preferences of some, it might be more immediately apparent that it is necessary to create ample opportunities for contestation.[59] Indeed, experts' claim to objectivity rests on the closure of the public debate. Therefore, when experts portray their decisions as a view from everywhere or nowhere, the need for sharing power appears as secondary.

To be sure, the IHRL project in many cases emphasizes the value of participation. David Kaye pays special attention to broadening the opportunities for more actors to participate in the governance of online speech, especially civil society from the Global South.[60] Evelyn Aswad, a leading scholar and member of the Board, calls for multi-stakeholder conversations.[61] The Board has often recommended Facebook conduct more extensive

53. Jasanoff, *supra* note 20, at 313 (describing a mode of building objectivity in regulatory science that she describes as a view from nowhere).

54. *Id*. at 315.

55. *Id*.

56. *See infra* Section II.A.

57. Daniel Walters, *The Administrative Agon: A Democratic Theory for a Conflictual Regulatory State*, 132 YALE L.J. (forthcoming 2022) (discussing the democratic foundations of the administrative state and arguing that what constitutes the common good is permanently contested).

58. *Id*.

59. *See* Nikolas Bowie, Comment, *Antidemocracy*, 135 HARV. L. REV. 160, 160 (2021) (defining democracy as the situation in which "everyone in the community, or *demos*, [may] share in exercising power, or *kratos*").

60. SR Report 2018, *supra* note 9, at para. 54.

61. Aswad, *supra* note 23, at 57.

stakeholder engagement.[62] However, as I argue in detail in Section IV, these calls do not challenge the role of IHRL as an account of the global public interest. Reasonable conflicts among diverse viewpoints might play a role in refining details, but do not constitute the core of the project.

Because judicial courts often aim at performing objectivity, many of the strategies I describe here should sound familiar to legal scholars.[63] Indeed, the Board replicates a judicial model in its quest to construct legal norms as exogenous principles it applies in each decision. The focus of the following sections will be on unveiling the deeply political interpretative work behind these principles and their application. Gaining a better understanding of the world that the project envisions, the tools it has to achieve it, and the intended and unintended effects of deploying those tools may lead to a more comprehensive evaluation of the IHRL project's potential contributions and risks.

Two clarifications are in order before proceeding.

First, this Article engages with how legal scholars, advocacy groups, and the Board justify and propose to implement IHRL as a project for online speech governance by corporate actors. The Article is, then, not a critique of IHRL more broadly. Indeed, in some cases, a disconnect exists between international law literature and how international law arguments are deployed in this specific context. Here I focus only on the latter and aim at making the disconnect explicit when relevant.[64]

Second, this Article is not about the intent, motivations, or state of consciousness of the actors promoting what I call the IHRL project for content moderation. When experts frame their proposals and decisions as mandated by international law or some other exogenous source, they might be aware that they are using legal language to cover a policy choice or might believe that their decision is in fact mandated by an external authoritative source, or might believe something else. Their state of mind is in no way part of this Article's inquiry.

II. Justifying the International Human Rights Law Project

IHRL is often invoked as the proper benchmark to evaluate content moderation without explaining why that would be appropriate.[65] In debates about online speech governance, it is often assumed that human rights law is

62. *See infra* Section IV.B.

63. Legal realism scholars mostly accept that legal reasoning conceals policy decisions. *See, e.g.,* Oliver Holmes, *The Path of the Law*, 10 Harv. L. Rev. 457 (1897); John Dewey, *Logical Method and Law*, 10 Cornell L.Q. 17 (1924); Karl Llewllyn, *Some Realism about Realism—Responding to Dean Pound*, 44 Harv. L. Rev. 1222 (1931).

64. *See infra* Section II.A.3, Section III.B.

65. *See infra* Section II.A.

synonymous with the public interest.[66] The attractive features of the framework are easy to see: its claim to universality and the high esteem in many circles. Thus, it could be useful to guide the private governance of speech on platforms that also intend to be global and have a great need for a legitimate set of rules to conduct their businesses.

On a closer look, however, it is not obvious that a framework designed for states would be appropriate for corporate actors. Accordingly, proponents of the IHRL project hint at various justifications for their proposal as well as modifications that they consider necessary to adjust the framework for the corporate environment. Examining these rationales serves two purposes. On the one hand, this section shows how the justifications most commonly offered share the quest for objective and apolitical principles to function as the foundations of speech governance. On the other hand, although these rationales are offered as objective reasons to use IHRL as the framework for content moderation, this section challenges their objectivity. It argues that these justifications are often internally inconsistent or depart widely from established principles of international law. In addition, I argue that the proposed adjustments to IHRL are closer to the construction of a new system of rules than to an exercise of translation.[67] As a consequence, even if IHRL were a universal framework with democratic buy-in, it is unclear that the IHRL project for content moderation would preserve those credentials.

I explore here five common rationales for asking social media companies to implement IHRL in their content policies: the UNGPs, a functional analogy between social media companies and states, the wide ratification of certain human rights conventions, the capacity of IHRL to rein in corporate power, and the fact that IHRL reflects globally shared values.

A. The Rationales

1. *A New Interpretation of the U.N. Guiding Principles*

Perhaps the most common justification for the IHRL project is the UNGPs. John Ruggie developed these principles during his tenure as the U.N. Secretary-General's Special Representative for Business and Human Rights. The instrument provides standards to guide the conduct of states and business enterprises. The UNGPs set the expectation that companies

66. *See* Douek, *supra* note 30, at 40 ("there has been almost no strong dissent from the proposition that IHRL should be adopted by companies as the basis for their rules"); Griffin, *supra* note 30, (describing human rights as "the rarely-questioned moral yardstick against which all platform practices and state regulation are measured"). Forceful opposing views to international human rights law as representative of the public interest do exist in the broader literature outside of the specific debates about content moderation. *See, e.g.,* Samuel Moyn, The Last Utopia: Human Rights in History (2010); David Kennedy, The Dark Sides of Virtue: Reassessing International Humanitarianism (2005).

67. *See* Sander, *supra* note 30, at 969, 971 (describing the IHRL project as a translation exercise).

will respect human rights wherever they conduct their operations.[68] On that basis, scholars interpret the UNGPs as calling on social media companies to align their policies with IHRL.[69] In that vein, the Board states in every decision that it applies international law per the UNGPs.[70] But the origins of the UNGPs show that this recent interpretation contradicts the history of the instrument as well as its drafter's understanding of its meaning, and it is hard to find support for this recent interpretation in the text of the UNGPs.[71]

To be sure, the UNGPs do provide guidelines that would help reduce the harm caused by online speech. According to these principles, companies should assess how their products impact people's human rights in the different markets where they operate.[72] Companies should take measures to stop any contribution to human rights violations.[73] This responsibility means that companies ought to take action regarding speech that international law prohibits.

The UNGPs, however, do not set the expectation that companies will align their content policies with IHRL. The UNGPs set the expectation that companies will respect human rights, but international law does not currently recognize a human right to speak on privately owned media without being subject to the medium's editorial policies. Neither does it establish a corresponding duty for media owners to host all speech that international law protects.[74] To the contrary, as Molly Land says, "users seek these platforms precisely because of the choices the companies make about the infor-

68. U.N. Special Representative of the Secretary-General, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, Principles 11-24, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011).

69. *See, e.g.,* Stefania Di Stefano, *The Facebook Oversight Board and the UN Guiding Principles on Business and Human Rights: A Missed Opportunity for Alignment?*, *in* Human Rights Responsibilities in the Digital Age 93 (Jonathan Andrew & Frédéric Bernard eds., 2021).

70. The Facebook Oversight Board has applied human rights standards per the UNGPs in all of its decisions. Since Facebook's adoption of its corporate human rights, the Facebook Oversight Board has also referenced this document as a basis for the application of international standards. *See Case decision 2021-003-FB-UA*, Oversight Bd., (Apr. 29, 2021), https://oversightboard.com/decision/FB-H6OZKDS3/ [https://perma.cc/VV4K-NMRH]*; see also Corporate Human Rights Policy*, Meta, https://about.fb.com/wp-content/uploads/2021/03/Facebooks-Corporate-Human-Rights-Policy.pdf [https://perma.cc/8FBK-XFUG] ("We are committed to respecting human rights as set out in the United Nations Guiding Principles on Business and Human Rights.").

71. Coincidentally, a recent report submitted by the U.N. High Commissioner for Human Rights to the Human Rights Council addressing the application of the UNGPs to the technology sector does not indicate that social media should align their content moderation policies to international human rights law. *See* OHCHR, *The Practical Application of the Guiding Principles on Business and Human Rights to the Activities of Technology Companies*, U.N. Doc. A/HRC/50/56 (Apr. 21, 2022).

72. U.N. Special Representative of the Secretary-General, *supra* note 68, Principle 13.

73. *Id.*, Principle 11. *See also* Allison-Hope et al., *A Human Rights-Based Approach to Content Governance*, BSR (Mar. 2, 2021), https://www.bsr.org/en/our-insights/blog-view/beyond-user-realdonaldtrump-human-rights-based-approach-content-governance [https://perma.cc/LGC9-ELD8].

74. *See* Matthias C. Kettemann & Anna Sophia Tiedeke, *Back up: can users sue platforms to reinstate deleted content?*, 9 Internet Pol'y Rev. 1 (2020) (showing that even in Germany, where courts have pioneered the development of intermediaries' duties to host certain types of content, platforms are allowed to exclude lawful expressions).

mation they deliver."[75] Accordingly, because excluding lawful content is an essential aspect of what social media companies do, IHRL does not fit well with how platforms operate, as Emily Laidlaw points out.[76]

In this context, it is unclear which existing human right the companies are called upon to respect when they are asked to regulate content in the same terms that international law binds state regulation of expression.[77] Indeed, the scholars who conclude that the UNGPs set the expectation that social media companies align their policies with IHRL are advancing an interpretation of this instrument that diverges from its original meaning.

Since the 1970s, the United Nations has tried to address the negative impact that corporations can have on human rights.[78] The direct antecedent of the UNGPs were the failed U.N. Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights 2003 (hereinafter "Draft U.N. Norms").[79] This initiative tried to create binding standards directed at transnational corporations implicated in abuses such as employing child labor, failing to provide safe working conditions, dumping toxic wastes, and disrupting the right to bargain collectively.[80] The Draft U.N. Norms attributed to business enterprises the same general obligations that states have to promote, secure the fulfillment of, respect, and protect human rights under the human rights treaties that states have ratified.[81]

The project was stalled following opposition from several states and the business sector.[82] The authors of the Draft U.N. Norms and the NGOs that participated in the drafting process believed that IHRL should apply directly to business enterprises.[83] The corporate sector and most states disagreed.[84] The U.N. Sub-Commission on the Promotion and Protection of

75. Land, *supra* note 40, at 292.

76. Emily B. Laidlaw, Regulating Speech in Cyberspace: Gatekeepers, Human Rights and Corporate Responsibility 292 (2015).

77. *See* SR Report 2018, *supra* note 9, ¶ 45.

78. *See generally* Surya Deva, *The UN Guiding Principles on Business and Human Rights and Its Predecessors: Progress at a Snail's Pace?*, *in* The Cambridge Companion to Business & Human Rights Law 145 (Ilias Bantekas & Michael Ashley Stein eds., 2021) (mapping the efforts of the United Nations to address the human rights duties of business enterprises).

79. Sub-Comm'n on the Promotion and Prot. of Hum. Rts., Comm'n on Hum. Rts., *Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with regard to Human Rights*, U.N. Doc E/CN.4/Sub.2/2003/12/Rev.2 (Aug. 26, 2003).

80. David Weissbrodt & Muria Kruger, *Norms on the Responsibilities of Transnational Corporations and Other Business Enterprises with Regard to Human Rights*, 97 Am. J. Int'l. L. 901 (2003) (detailing the history, purpose and content of the norms).

81. Sub-Comm'n on the Promotion and Prot. of Hum. Rts., Comm'n on Human Rights, *supra* note 79.

82. *See* Deva, *supra* note 78, at 155; John G. Ruggie, *Incorporating Human Rights: Lessons Learned, and Next Steps*, *in* Business and Human Rights: From Principles to Practice 64, 65 (Dorotheé Baumann-Pauly & Justine Nolan eds., 2016).

83. Weissbrodt & Kruger, *supra* note 80, at 906.

84. John G. Ruggie, Just Business: Multinational Corporations and Human Rights 47–55 (2013).

Human Rights approved the norms in August 2003, but the Human Rights Council (then Commission) refused to endorse them.[85]

In that context, the Human Rights Council created John Ruggie's mandate in 2005. The original mandate of the Special Representative of the Secretary-General for Business and Human Rights was to "identify and clarify standards of corporate responsibility and accountability" for businesses concerning human rights and to elaborate the role of states in effectively regulating corporations.[86] Ruggie "saw no reason to replicate the debate" that had bogged down the U.N. Draft Norms.[87] Thus, he introduced some modifications to the previous framework to gain support from states and corporations.[88]

In Ruggie's vision, the U.N. Draft Norms "had serious foundational flaws, such as intermingling state and corporate obligations while providing no boundaries for the latter."[89] Accordingly, in the UNGPs, he distinguished states' obligations from corporations' responsibilities. Unlike the U.N. Draft Norms, which would have imposed legal obligations on corporations, the UNGPs set expectations that enterprises should meet voluntarily. However, the distinction was not merely that states were bound by international law and businesses should follow it voluntarily. Instead, Ruggie introduced a substantive distinction between the content of states' duties and businesses' responsibilities.[90]

According to Ruggie's framework, companies should look at international treaties not as a source of rules that apply to them but as an enumeration of recognized rights.[91] Thanks to this differentiation, even states that had not ratified core U.N. human rights treaties endorsed the UNGPs. Notably, China and the United States ratified them even though China has not ratified the International Covenant on Civil and Political Rights, and the United States has not ratified the International Covenant on Economic, Social and Cultural Rights.[92]

Corporations' main responsibility is to "avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved."[93] Some emblematic cases help to clarify the com-

85. Comm'n on Hum. Rts., *Report on the Sixtieth Session*, ¶41, U.N. Doc. E/CN.4/2004/L.73/Rev.1 (2004).

86. Office of the High Comm'r for Hum. Rts., *Secretary-General to appoint a special representative on the issue of human rights and transnational corporations and other business enterprises*, U.N. Doc E/CN.4/RES/2005/69 (Apr. 20, 2005).

87. Ruggie, *supra* note 82, at 65.

88. *Id.*

89. John Gerard Ruggie, *The Social Construction of the UN Guiding Principles on Business and Human Rights*, *in* Research Handbook on Human Rights and Business 63, 71 (Surya Deva & David Birchall eds., 2020).

90. Ruggie, *supra* note 84, at 81-127.

91. Ruggie, *supra* note 89, at 71.

92. Ruggie, *supra* note 82, at 65.

93. U.N. Special Representative of the Secretary-General, *supra* note 68, Principle 11.

panies' responsibilities. By 1990, Nike had outsourced its entire production. Its overseas sourcing factories employed over 24,000 workers in Asia.[94] Workers claimed problematic practices such as not being allowed to leave the premises except on Sundays and, even then, needing authorization from management.[95] *Life* magazine published a photograph of a twelve-year-old boy stitching soccer balls.[96] Nike's initial defense was that they did not own the factories.[97] In response, the UNGPs set the expectation that companies should conduct human rights due diligence to learn how their operations may contribute to or be involved in human rights abuses, and address this. For example, in this case, Nike would have been expected to recognize that child labor and other abusive practices infringe on human rights that international treaties identify and to avoid engaging in those abuses.[98]

Facebook's involvement in the genocide in Myanmar also illustrates corporate failure to meet this expectation. The report of the independent international fact-finding mission on Myanmar found that "Facebook has been a useful instrument for those seeking to spread hate, in a context, where, for most users, Facebook is the Internet."[99] The report stressed that Facebook's inability to provide country-specific data about the spread of hate speech on its platforms made it difficult to assess the adequacy of Facebook's response to the situation.[100] This type of risk assessment is the kind of due diligence that, under the UNGPs, companies are expected to undertake in order to evaluate and prevent their participation in human rights violations.

The IHRL project for content moderation interprets the UNGPs in a new fashion. Under its original meaning, social media companies are expected not to infringe on individuals' rights to freedom of expression (and any other rights) to the extent that international law already recognizes those rights. The IHRL project proposes that companies ensure a new right to freedom of expression on privately-owned social media companies.

The project collapses the substance of states' duties and businesses' responsibilities. Even though they propose adaptations,[101] IHRL advocates call on companies to look at the text of international treaties and the interpretations that the U.N. Human Rights Committee and U.N. Special Rapporteurs have made of them as a set of rules that companies should apply on

94. Jennifer Burns & Debora Spar, *Hitting the Wall: Nike and International Labor Practices, Case 9-700-047,* Harv. Bus. School Case Collection (2002).

95. *Id.*

96. Sydney Schanberg, *Six Cents an Hour*, Life (Mar. 28, 1996).

97. *Id.*

98. *See* Ruggie, *supra* note 84, at 3-6, 17, 69.

99. Human Rights Council, *Report of the independent international fact-finding mission on Myanmar*, ¶ 74, U.N. Doc. A/HRC/39/64 (Sept. 12, 2018).

100. *Id.*

101. *See, e.g.,* Benesch, *infra* note 127.

their platforms.[102] This is the approach adopted by the U.N. Draft Norms, which the Human Rights Council rejected and from which the UNGPs explicitly diverged.[103]

At the same time, this new interpretation of the UNGPs alters the interplay between states' duties and corporate responsibilities. Arguing that every content moderation decision implies an IHRL issue means that states have a corresponding duty to regulate each instance of content moderation.[104] The IHRL project's interpretation of the UNGPs creates a human right to freedom of expression on social media platforms vis-à-vis the owners of such platforms.[105] If individuals have that right, states have a corresponding duty to protect it.[106] However, there is little support, if any, for the idea that states have a duty to protect the exercise of all IHRL-protected expression on private platforms.[107] Indeed, no state requires companies to host all speech that international law protects.[108] The IHRL project for content moderation asks from corporations what international law does not require states to protect.

Another innovation refers to the different levels of responsibility between large and small firms. According to the UNGPs, all firms are expected to respect all internationally recognized rights.[109] Does that mean that all companies, including the New York Times or the Mandarin Oriental Hotel Group,[110] need to follow Article 19 of the ICCPR when deciding what speech they will host? This interpretation of the UNGPs would be absurd and hard to reconcile with the independence of the media and the protection of their editorial freedom.[111] In response, IHRL advocates sometimes rely on Principle 14 of the UNGPs to distinguish among the responsibilities of prominent social media companies and other business enterprises.

Principle 14 of the UNGPs establishes that the corporate responsibility to respect human rights applies to all enterprises. However, the means through

102. *See, e.g.,* Aswad, *supra* note 23; Parmar, *supra* note 10; *Side-stepping rights: Regulating speech by contract*, ARTICLE 19 (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB.pdf [https://perma.cc/ZB6R-LCLB] (hereinafter ("Side-stepping rights").

103. *See supra* n. 57-68 and accompanying text.

104. *See* Land, *supra* note 40, at 292 (making a similar point stating that "it cannot be the case that every content moderation decision made by every digital platform should be subject to human rights scrutiny.").

105. *See supra* Section II.A.1.

106. U.N. Special Representative of the Secretary-General, *supra* note 68, Principle 1.

107. *See, e.g.,* Balkin, *infra* note 169, at 2025 (explaining why the best alternative to the current autocracy is not imposing the duty to carry all lawful speech).

108. *See* Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech*, 1902 Hoover Inst. Aegis Paper Series, 12 (2019) (surveying laws that impose the legal duty on platforms to host speech, none of which includes the duty to host all lawful speech).

109. Ruggie, *supra* note 84, at 79.

110. The Mandarin Oriental group is one of 35 hotel companies and hundreds of other companies that have adopted a human rights policy and committed to respect human rights. *See Companies*, BUS. & HUM. RTS. RES. CTR., https://www.business-humanrights.org/en/companies/ [https://perma.cc/X68U-9SWS] (tracking companies' human rights policies).

111. *See* General Comment 34, *supra* note 34, ¶ 16.

which enterprises meet that responsibility may vary according to their size, sector, operational context, ownership, and structure.[112] Some authors who support the IHRL project indicate that this principle gives companies flexibility in how they apply human rights norms to their operations.[113] Accordingly, Principle 14 could justify that smaller platforms with specific purposes, more limited resources to moderate content, or specific targeted audiences may depart from international guidance to regulate speech.

This distinction in the UNGPs was originally intended to recognize that some firms have more extensive resources to assert influence. Opportunities to exert more influence come with greater responsibility for protecting human rights. Going back to the Nike example, a much smaller business that outsourced part of its production may not have the resources to audit in depth the whole supply chain it takes part in, or may not have the necessary influence to bring about change. In the words of David Weissbrodt and Muria Kruger, "[t]his nuanced approach does not lower the standards for any business; it simply ensures that those with greater power and influence will also have greater responsibilities."[114]

Companies may devote different amounts of resources to their human rights due diligence, but the content of the rights that companies must respect is always the same.[115] According to the interpretation advanced by the IHRL project, however, large and small companies have widely different substantive responsibilities. Large companies should adopt international law as their own internal rules. In contrast, small companies may adopt rules that directly contradict international standards for speech regulation.

Finally, a note on the project's ambivalence between asking companies to implement IHRL and calling on companies only to take inspiration from IHRL.[116] It could be that the UNGPs require that companies will implement international law, but not as if they were a state.[117] Perhaps the UNGPs do not set the expectation that companies follow IHRL, but do set the expectation that companies ground their policies in human rights or use human rights as "an overall framework for decision-making and action[.]"[118] For example, perhaps companies need to balance the right to freedom of expression and the right to equality, but the rule they reach as a result of that balancing exercise might contradict international law as in the Black Pete case. As I discuss in detail in the next subsection, it is unclear why the new rule would be called IHRL. The value of labeling these new standards as IHRL might reside in preserving the legitimacy credentials of

112. U.N. Special Representative of the Secretary-General, *supra* note 68, Principle 14.

113. *See, e.g.,* Benesch, *infra* note 127, at 95; Aswad, *supra* note 23, at 39.

114. Weissbrodt & Kruger, *supra* note 80, at 911. *See also* Ruggie, *supra* note 84, at 101, 114.

115. *Id.*

116. *See supra* Section I.

117. *See, e.g.,* Land, *supra* note 40, at 305-306 (offering a thoughtful description of how private companies might follow Article 19 of the ICCPR in order to meet the expectations set by the UNGPs).

118. Allison-Hope et al., *supra* note 73.

IHRL, even if the connection between the new rule and IHRL is thin at best.

Ruggie's interpretation of the UNGPs is not necessarily the best one. It may be that IHRL advocates have created a new way of reading the principles that advances worthy policy objectives like expanding users' rights to freedom of expression vis-à-vis private actors. My aim is to show both the urge to frame policy positions as mandated by legal documents as well as the intensive intellectual work behind the purportedly objective principles under which experts govern online speech.

2. *Because social media companies are like states, but they are not like states*

Scholars raise the concern that companies that control giant social media platforms are displacing governments as the main speech regulators. In the words of Richard Ashby Wilson and Molly Land:

> Governments are no longer the primary regulators of speech. Their regulatory capacity has been far outstripped by some of the largest companies in the world . . ., which together regulate the speech of 3.7 billion active social media users. . . . In a reversal of the historic roles, private corporations have even become the de facto regulators of government speech[.][119]

Scholars conceptualize what giant platforms do as state functions. Nadine Strossen states, "the Platforms wield censorial power of a magnitude that in the past only governments have exercised."[120] Julie Cohen says, "[d]ominant platforms' role in the international legal order increasingly resembles that of sovereign states."[121] Daphne Keller agrees, stating that, "platforms can take on and replace traditional state functions, operating the modern equivalent of the public square or the post office, without assuming state responsibilities."[122]

In most aspects, platforms are nothing like states. Platforms do not collect taxes, manage prisons, hold elections, etc.[123] The role of technology companies is hardly unprecedented. Other highly concentrated media industries have controlled the public sphere in the past.[124] However, the analogy can be a useful rhetorical device to highlight the immense power platforms wield over public discourse and collective affairs.

---

119. Richard Ashby Wilson & Molly Land, *Hate Speech on Social Media: Content Moderation in Context*, 52 Conn. L. Rev. 1, 5 (2021).

120. Nadine Strossen, *United Nations Free Speech Standards as the Global Benchmark for Online Platforms' Hate Speech Policies*, 29 Mich. State Int'l L. Rev. 307, 324-5 (2021).

121. Julie Cohen, *Law for the Platform Economy*, 51 U.C. Davis L. Rev. 133, 199 (2017).

122. Keller, *supra* note 108, at 2-3.

123. *See* Benesch, *infra* note 127 (highlighting the obvious but fundamental fact that "Facebook is not a country.").

124. Paul Starr, The Creation of the Media: Political Origins of Modern Communications (2004).

Describing platforms as states can have profound normative implications. For example, one could imagine users as citizens or political subjects with an interest in participating in policy making. The IHRL project emphasizes a different implication of this analogy: companies should meet the requirements that states must meet to restrict expression under human rights treaties. This emphasis on adhering to existing legal frameworks for states (as opposed to other possible derivations from the analogy between platforms and states) is also more compatible with the functioning of the Board. The Board operates under a judicial model. That background helps explain the gravitational force attracting the Board to a system of legal norms.

Because platforms are different from states, scholars propose adaptations to the international law framework. Proponents of the IHRL project acknowledge that the first two requirements set out by Article 19 of the ICCPR—legality and legitimacy—need to be "translated."[125]

Regarding the legitimacy prong, supporters of the IHRL project struggle to define what aims content moderation could legitimately pursue. Article 19 of the ICCPR enumerates legitimate aims for governmental restrictions on speech: respect for the rights or reputations of others, the protection of national security, and public order, public health, or morals.[126] However, these aims are both over-inclusive and under-inclusive for platforms. On the one hand, Susan Benesch argues that firms are not well positioned to determine, for example, national security goals.[127] On the other hand, Evelyn Aswad asks if it would be legitimate for companies to pursue the aim of maximizing profit.[128] Aswad believes IHRL does not provide a conclusive answer here. In her version of the IHRL project, multi-stakeholder conversations should define the legitimate aims for corporate IHRL. In practice though, this specific discussion about commercial aims may not be that relevant. Companies will usually find it easy to develop a public-interest rationale to restrict speech without invoking business reasons.[129]

Suppose the IHRL project concedes that the aims defined by Article 19 are ill-suited for private speech regulation and accepts, instead, any public-interest purpose. In that case, the IHRL project would probably approve any aim that companies enunciate as legitimate. In fact, in all decisions to date, the Board has found that Facebook's community standards have a legitimate aim.

125. *See* Sander, *supra* note 44.

126. ICCPR, *supra* note 36, art. 19(3).

127. Susan Benesch, *But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies*, 38 YALE J. ON REG. BULL. 86, 106 (2020).

128. Aswad, *supra* note 23.

129. *See, e.g., Facebook Community Standards*, *supra* note 29 (including a policy rationale for each community standard).

The requirement of legality has also been partially eroded. This prong demands that restrictions to freedom of expression are "provided by law."[130] That is, there must be a legal basis for restrictions. In other words, only bodies authorized to make law may impose such limits.[131] The legality prong has also been interpreted to require that restrictions be subject to public comment and that independent judicial officials oversee their implementation.[132]

The Human Rights Committee has argued that the requirement refers not only to the body authorizing the restriction but also to its clarity and precision. A law must be formulated with sufficient accuracy so that individuals can regulate their conduct accordingly.[133] This aspect of the requirement has been very generative for the IHRL project. In the context of content moderation, the provision is understood to mean that moderation practices must be clear and transparent.[134]

The first part of the legality requirement, which refers to the entity that may restrict individuals' freedom of expression, is impossible for companies, as they exist today, to meet. Again, seeing companies as states could be a first step towards imagining participatory institutions. Instead, the legality part of the test has been narrowed down to the transparency requirement.

Both transformations risk leaving too little of IHRL standing. What is left is a test that says that restrictions on speech should be necessary to achieve a legitimate goal. Without any guidance about which aims are permissible or who should issue the restrictions, what is left is a common-sense idea that rules should reasonably serve a reasonable purpose.

Overall, IHRL is thought of as a constraint on state power that can also constrain corporate power. On closer examination, IHRL requires the existence of state institutions to function effectively. Corporations do not have internal systems authorized to pass laws in the terms of Article 19 of the ICCPR, nor do they have institutions capable of making decisions about national security or the adequate balance between human rights. IHRL cannot make up for the lack of these structures; it can only demand that when these institutional actors are available, they must regulate speech in accordance to Article 19 of the ICCPR and any other relevant provisions. The

130. ICCPR, *supra* note 36, art. 19(3). *See also* Gilad Abiri & Sebastián Guidi, *From a Network to a Dilemma: The Legitimacy of Social Media*, STAN. TECH. L. REV. (forthcoming 2023) (manuscript at 32-3) (available at https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4230635) (making a similar point).

131. General Comment 34, *supra* note 34, ¶ 24; SR Report 2016, *supra* note 34, ¶ 12.

132. SR Report 2016, *supra* note 34, ¶¶ 12, 13; SR Report 2018, *supra* note 9, ¶ 7; David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Promotion and Prot. of the Right to Freedom of Op. and Expression*, ¶ 6(a), U.N. Doc. A/74/486 (Oct. 2019) (hereinafter "SR Report 2019").

133. General Comment 34, *supra* note 34, ¶ 25.

134. *See* Sander, *supra* note 30, at 971; Benesch, *supra* note 127, at 103; Aswad, *supra* note 23, at 46. However, calls for transparency in content moderation are sound, long-standing, and widely accepted. It is unclear whether the IHRL project has added any force or meaning to these calls. *See, e.g.,* THE SANTA CLARA PRINCIPLES, *supra* note 8.

tension then lies in the fact that even if companies do what states ought to do, corporations do not have the institutions that are critical pieces to operate the IHRL machinery.

3. *Because IHRL is consented to by all states, but state consent does not matter*

Nearly everyone seems to be well aware that no universal, or even local, agreement on how to regulate speech exists. Even international law scholars deeply question the universality of international law.[135] However, when justifying the IHRL project, advocates remind us that IHRL is the only globally-adopted framework. As Evelyn Aswad says, "[c]ompanies need not recreate the wheel in developing speech norms that have worldwide legitimacy if they base their content moderation policies on international human rights standards."[136]

In some accounts of the IHRL project, its claim to universality comes from state consent, particularly the vast adoption of U.N. treaties.[137] As former ACLU president Nadine Strossen puts it, "almost every single country in the world is a party to the ICCPR and the ICERD."[138] Proponents of the IHRL project also invoke the wide acceptance of the UNGPs as one of the main reasons to acknowledge their legitimacy. Strossen and Aswad emphasize that the Human Rights Council endorsed the UNGPs unanimously.[139]

From these statements, one could assume that states would be a good source for identifying a community's values or agreements. However, IHRL advocates regard states with deep suspicion. When Aswad analyzes a proposal to renegotiate and clarify some international treaties, she rejects that possibility (for understandable reasons): "an international negotiation to regulate speech on platforms, including content moderation, is undesirable because it would no doubt be dominated by powerful countries with weak records on freedom of expression that would seek to roll back international speech protections."[140] Instead of opening up avenues for states to deliberate, Aswad proposes that we rely on experts' interpretations that emphasize the interpretations of U.N. treaties that broaden protections for freedom of expression.[141]

---

135. *See* ANTHEA ROBERTS, IS INTERNATIONAL LAW INTERNATIONAL? (2017).

136. Aswad, *supra* note 23, at 65.

137. *See, e.g., id.* at 35.

138. Strossen, *supra* note 120, at 329. *See also* Benesch, *supra* note 127, at 89 ("No source of rules for speech regulation is as widely known or formally adopted as international human-rights law. That law's most relevant instrument, the International Covenant on Civil and Political Rights (ICCPR), has been ratified by nearly ninety percent of the countries in the world.").

139. Strossen, *supra* note 120, at 355-56; Aswad, *supra* note 23, at 38.

140. Aswad, *supra* note 23, at 61.

141. Evelyn Aswad, *To Protect Freedom of Expression, Why Not Steal Victory from the Jaws of Defeat?*, 77 WASH. & LEE L. REV. 609, 648-649 (2020).

When IHRL advocates discuss the divergences between U.N. and regional treaties, they disregard state consent as an essential factor in determining which rules should govern speech. Interestingly, they acknowledge that all regional systems diverge to various extents from U.N. standards.[142] If state consent were the source of IHRL's universality, these conflicts of norms should be read as an opportunity to determine which commitment more accurately reflects the state's intent, values, or norms. However, promoters of the project often assert that companies should follow only U.N. guidance. For example, Aswad states that "*regional* human rights instruments (and monitoring bodies) are not *international* human rights instruments (and monitoring bodies)."[143] Global expectations, she explains, are reflected in international instruments, not regional ones.[144]

If it is the wide ratification of international treaties that makes these norms a reflection of global values, it is unclear why the ratification of contradictory treaties would be dismissed as irrelevant. A way out of this tension is to see human rights' universality as prior to state consent and not dependent on it. The wide ratification of human rights would only confirm their universal status, rather than constitute it. In that case, IHRL may be universal because it reflects global ethical principles that align with the public interest. The next two subsections discuss this alternative claim to universality.

4. *Because IHRL constrains corporate power, but its indeterminacy is a feature*

The IHRL project's central promise is that it can align corporate governance with the public interest. However, IHRL is also justified as an appropriate framework because it does not mandate specific policies.

IHRL could put the public interest at the center of speech regulation if it reflected clear ways of delimiting the scope of conflicting rights that are either globally shared or ethically correct. In response to concerns that IHRL is too vague or internally contradictory,[145] some scholars argue that IHRL provides precise enough standards on many valuable points. Evelyn Aswad and organizations such as Article 19 have unpacked what IHRL demands

142. *See, e.g.,* Evelyn Aswad & David Kaye, *Convergence & Conflict: Reflections on Global and Regional Human Rights Standards on Hate Speech*, 20, NW. J. HUM. RTS., 165 (2022).

143. Aswad, *supra* note 23, at 44.

144. *Id*. *See infra* Section III.B. (addressing how IHRL advocates try to iron out this contradiction between state consent as the source of the legitimacy of IHRL and IHRL's indifference toward state consent).

145. *See generally* Amal Clooney & Philippa Webb, *The Right to Insult in International Law*, 48 COLUM. HUM. RTS. L. REV. 1 (2017); Douek, *supra* note 30, at 37; 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 [https://perma.cc/S7AE-N3AT].

from corporations.[146] David Kaye's report on hate speech as Special Rapporteur offered clear principles that companies should follow.[147]

Despite these efforts to unpack IHRL into detailed rules, IHRL has areas of indeterminacy even in its most granular versions, and its most well-established principles admit exceptions. Hate speech bans provide a good example. Kaye explains that bans on statements that deny well-established historical atrocities such as the Holocaust or the Armenian genocide are incompatible with IHRL. According to the Special Rapporteur's 2018 report: "offensive interpretation of a religious tenet or historical event . . . is not to be silenced under Article 20 (or any other provision of human rights law)."[148] The same report states that "[l]aws that penalize the expression of opinions about historical facts are incompatible with Article 19 of the Covenant, calling into question laws that criminalize the denial of the Holocaust and other atrocities and similar laws, which are often justified through references to hate speech."[149] Kaye immediately adds that "the application of any such restriction under IHRL should involve the evaluation of the six factors noted in the Rabat Plan of Action."[150] Therefore, even prohibitions that seem clear enough to guide companies on highly controversial questions admit exceptions under vague circumstances.

The technical singularities of social media increase the indeterminacy of IHRL. The organization Article 19 explains that most corporate policies on hate speech are too broad to be compatible with IHRL because they ban expressions that do not incite violence or illegal action.[151] However, Kaye clarifies that "[a]cross a range of ills that may have a more pronounced impact in digital space than they might offline . . . human rights law would not deprive companies of tools."[152] Rather, "it would offer a globally recognized framework for designing those tools and a common vocabulary for explaining their nature, purpose, and application to users and States."[153] While some see IHRL as limiting bans on offensive content that does not incite illegal action, Kaye tells us that these bans could be acceptable in the online context.

---

146. *See Side-stepping rights*, *supra* note 102.

147. SR Report 2018, *supra* note 9; SR Report 2019, *supra* note 132. *See also* Michael O'Flaherty, *International Covenant on Civil and Political Rights: interpreting freedom of expression and information standards for the present and the future*, *in* THE UNITED NATIONS AND FREEDOM OF EXPRESSION AND INFORMATION 55, 82 (Tarlach McGonagle & Yvonne Donders eds., 2015).

148. SR Report 2018 *supra* note 9, ¶ 10.

149. *Id.*, ¶ 22. *See also* Aswad & Kaye, *supra* note 142.

150. SR Report 2018 *supra* note 9, ¶ 22; Human Rights Council, *Ann. Rep. of the U.N. High Comm'r for Hum. Rts. Addendum*, U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013) (outlining a six-part threshold test taking into account (1) the social and political context, (2) status of the speaker, (3) intent to incite the audience against a target group, (4) content and form of the speech, (5) extent of its dissemination and (6) likelihood of harm, including imminence).

151. *Side-stepping rights*, *supra* note 102, at 19.

152. SR Report 2018, *supra* note 9, ¶ 43.

153. *Id.*

At the same time, IHRL's indeterminacy is praised as a positive feature. Because IHRL is, like any legal system, indeterminate, its adoption does not impose answers. Instead, it offers a margin of maneuver for companies to choose their house rules[154] and a process for actors to engage in conversation and to reach those answers.[155] However, it is unclear how IHRL would facilitate conversations between different actors. If actors are not included in the discussion, it is more likely due to power and resource differentials rather than the need for a common language to understand each other.[156]

The two sides of the debate emphasize potentially positive aspects of IHRL: it allows a diversity of regulatory approaches and it provides some parameters for acceptable rules. However, a tension exists between arguing that IHRL offers global standards that companies can voluntarily adhere to and arguing that IHRL's main advantage is that it can create the necessary conditions for a multiplicity of actors to decide what those standards should be.

5. *Because IHRL is a shared language and because IHRL is like the First Amendment*

IHRL is portrayed as "a shared language" or least a common denominator.[157] It is thought to provide a baseline for (unidentified) actors to be in conversation and reach more granular agreements with each other.

At the same time, U.S. scholars argue that calling on companies to adopt U.N. standards is the most promising avenue to bring content moderation closer to First Amendment doctrine.[158] They believe this strategy is auspicious because U.N. sources would be more readily accepted globally than the First Amendment doctrine, even though the two share many vital elements.[159]

Similarly, as previously discussed, when U.N. authoritative interpretations conflict with decisions originating from European bodies, proponents of the IHRL project assert that U.N. solutions should prevail in the context of content moderation.[160] It is problematic to claim that a legal framework is open-ended, and either reflects or invites the search of collective answers, while arguing that U.N. solutions should always preempt regional or local preferences.

---

154. SR Report 2019, *supra* note 132, ¶¶ 43, 48; ARTICLE 19, *supra* note 43, at 13.

155. Sander, *supra* note 30, at 967-68.

156. Blayne Haggart & Clara Iglesias Keller, *Democratic legitimacy in global platform governance*, 45 TELECOMM. POL'Y 1, 11-12 (2021); David Kennedy, *The International Human Rights Movement: Part of the Problem?*, 15 HARV. HUM. RTS. J. 101, 109 (2002).

157. Michael Ignatieff, *Human Rights as Idolatry*, *in* TANNER LECTURES ON HUMAN VALUES 349 (2000) (articulating the value of human rights as a "shared vocabulary" in a broader context).

158. Strossen, *supra* note 120, at 333 ("Notwithstanding widespread assumptions about the exceptionally speech-protective nature of U.S. free speech law, careful comparison of the U.N. approach to that of the U.S. demonstrates that the two share more key elements than has generally been recognized.").

159. *Id.*

160. *See, e.g.,* Aswad, *supra* note 23, at 57.

B. Objective Justifications

These tensions, contradictions, and innovative interpretations show that the invoked reasons for corporations to apply international law are not objective. However, by portraying the reasons as such, it is possible to justify why the global society should perceive the governance of online speech by corporate actors and bodies of experts as advancing the public interest. Human rights experts' democratic buy-in allegedly comes from the framework that constrains them and that they implement. The decision to use this specific framework, in turn, must itself be a reflection of the common good or some general consensus. Each of the justifications examined above reflects this kind of discourse: experts shall apply IHRL because an exogenous instrument like the UNGPs mandates this application or because IHRL has been collectively agreed on as the appropriate framework to govern speech on social media.

The UNGPs, through hard work, have been reconstructed as an exogenous instrument, already endorsed unanimously by the Human Rights Council, that recommends that corporations align content policies with IHRL. In all the other justifications, the premise is that the "global society" already knows what rules should regulate speech.

In that sense, in the second justification (platforms are like states and we already know how states ought to restrict speech), the debate shifts from how society wants companies to moderate content to how the standards we supposedly already recognize as appropriate to govern speech can be operationalized. That is, this rationale is premised on the view that a broad agreement exists about how speech should be governed. Accordingly, what is needed is an adjustment to a new context that experts can carry out.

State consent is another expression of the ideal of objectivity: international law expresses what everyone—or every state—has agreed on. In this case, objectivity draws not on the rightfulness of the principles but on every state's participation. When IHRL advocates reference states' consent, they imagine that the participatory procedure has already occurred.[161] What is left are standards for speech regulation that can be implemented in different contexts. In the last two rationales, the ideal of objectivity can also be traced to an imagined consensus among global constituencies. IHRL is described as a shared language among societies globally.

Not all the reasons supporting the IHRL project, however, are premised on this ideal. When advocates say that IHRL is a good framework because it resembles the core principles of First Amendment jurisprudence,[162] these advocates champion their normative preferences. Even in this case, advocates

161. Aswad, *supra* note 141 (arguing that this participatory procedure should not be reopened to negotiate more granular rules applicable to social media platforms because the outcome could be less protective of freedom of expression than the current interpretations made by U.N. agencies).

162. *See supra* Section I.D.

believe that their normative preferences will be more convincing if framed as neutral principles. When Nadine Strossen compares the First Amendment to U.N. treaties, she emphasizes that the two "share more key elements than has generally been recognized."[163] However, Strossen is right to believe that it is strategic to frame advocacy efforts in terms of U.N. rules because they have a claim to universality that the First Amendment lacks.

Overall, to the extent that the IHRL project is concerned with making online speech governance more aligned with the public interest, the project imagines a governance model where a group of people implement a set of principles that reflect the interests of the global society. This section examined how advocates justify or imagine the availability of that set of principles. The next section analyses how experts and especially the Board perform as objective executors of those principles.

III. Deploying the IHRL Project

This section explores how decisionmakers, and especially the Board, navigate conflicts of norms and areas of indeterminacy within international law while preserving the project's claim to objectivity. When IHRL does not yield a specific outcome or experts are faced with multiple possible solutions, experts invoke interpretative theories, technical facts, or other strategies to conceal their policy choices. Decisions are justified as a logical deduction from higher principles or as the necessary consequence of objective facts. This section looks at four tools that the IHRL project has so far developed for expert governance, while paying close attention to the Board's first decisions.

A. IHRL as Self-Evidently Good

How is it that, despite so much disagreement about how platforms should moderate content, so many people believe that their views about online speech are based on human rights?[164] It may be that because this concept is so ill-defined, it can justify many policy proposals and decisions. All sides of any debate about restrictions on speech are defending a human right. Indeed, most interests usually invoked around speech adjudication can find an anchor in a human right: freedom of expression, safety, dignity, non-discrimination, equality.

Because of that open texture, "human rights" can be used as a synonym for the common good. A policy report by the organization Article 19 offers a clear example. This report describes the steps social media companies should take to comply with international freedom of expression standards.[165] When

163. Strossen, *supra* note 120, at 333.
164. *See supra* notes 21-26 and accompanying text.
165. *Side-stepping rights*, *supra* note 102.

analyzing platforms' internal hate speech rules, the report finds that platforms' rules usually restrict speech that IHRL protects.[166] According to the report, companies do not offer robust speech protections for commercial reasons. The authors hypothesize that these rules enable platforms to grow their user base and accommodate advertisers' interests.[167] Overall, the report assumes that there are two options: rules driven by business interests and rules that align with IHRL. At no point does the report question the idea that IHRL and the public interest are equivalent.

The comparison between conversations about platforms adopting First Amendment doctrines and IHRL as a default illuminates how IHRL is discussed as undoubtedly good. Most scholars who support First Amendment doctrine as a generally appropriate normative framework to govern speech are resistant to asking companies to apply it as a default rule for content moderation.[168] Scholars acknowledge that if platforms hosted all First Amendment–protected speech, platforms would be worse for almost all users.[169] However, similar concerns have not been raised by IHRL supporters, even though if social media were to allow all the speech that IHRL protects, users and companies would also have to tolerate all types of undesirable speech.[170] IHRL has both a non-contestable character and a higher level of indeterminacy that seems to incentivize people to accept it as a framework and work out exceptions rather than rejecting the whole framework because it does not perfectly track their normative preferences.

Because of its simultaneously undefined and uncontestable character, human rights language can be invoked to justify a decision without many further explanations. If someone proposes that platforms should adopt a certain policy because it aligns with the First Amendment, it is very likely that other advocates will raise serious objections to that justification. If someone, however, argues in favor of a rule because it follows human rights law, they are unlikely to be challenged on the basis that their framework is inadequate, even if they still have to persuade others about the merits of the specific rule.

166. *Id.* at 16.

167. *Id.*

168. *See* Dvoskin, *supra* note 26.

169. *See, e.g.,* Jack Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2025 (2018) ("[T]he best alternative to this autocracy is not the imposition of First Amendment doctrines by analogy to the public forum or the company town."); Keller, *supra* note 108, at 13; Jonathan Peters, *The Sovereigns of Cyberspace and State Action: The First Amendment's Application-Or Lack Thereof-To Third-Party Platforms*, 32 BERKELEY TECH. L.J. 989 (2018); Rebecca Tushnet, P*ower Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986, 988 (2008); Christopher Yoo, *Free Speech and the Myth of the Internet as an Unintermediated Experience*, 78 GEO. WASH. L. REV. 697, 700 (2010).

170. *See* Keller, *supra* note 108.

B. The Relationship between Global and Regional Norms

A recurrent problem that IHRL supporters face is the tension between the U.N. system and regional systems of human rights. Discussing the feasibility of the IHRL project, scholars Susan Benesch and Barrie Sander emphasize that human rights law can be confusing and inconsistent.[171] In particular, Evelyn Douek points out that regional treaties sometimes contradict each other or U.N.-level instruments.[172]

IHRL advocates acknowledge these conflicts. David Kaye mentions that the European Court of Human Rights has adopted a deferential approach toward states that ban blasphemy or criminalize genocide denial "in contrast to trends observed at the global level."[173] Aswad adds that most regional systems have points of tension with U.N. treaties.[174] The Cairo Declaration on Human Rights in Islam states that everyone shall have a right to express their opinions freely "in such manner as would not be contrary to the principles of the Shari'ah."[175] The Association of Southeast Asian Nations' human rights declaration limits freedom of expression in ways that are inconsistent with the Universal Declaration of Human Rights.[176] Recently, Aswad and Kaye have analyzed in depth the convergences and divergences of regional human rights systems and U.N. norms regarding hate speech standards and have concluded that regional norms often diverge from U.N. standards.[177]

Human rights tribunals often navigate these conflicts through jurisdictional rules. For example, even though international tribunals may be in dialogue with each other, the Human Rights Committee is tasked with interpreting the ICCPR while the Inter-American Court first looks at the American Convention on Human Rights.[178] This is also true for other kinds of tribunals: World Trade Organization tribunals normally give priority to WTO agreements, while human rights norms take precedence before human rights courts.[179] Similarly, national authorities might give more importance

171. Benesch, *supra* note 127, at 91; Sander, *supra* note 30, at 977.

172. Douek, *supra* note 30. *See also* Dvoskin, *supra* note 145.

173. SR Report 2019, *supra* note 132, ¶ 26 (referring to the U.N. level).

174. Aswad, *supra* note 136, at 634, n.94.

175. Org. of Islamic Cooperation, *The Cairo Declaration on Human Rights in Islam*, Annex to Res. No. 49/19-P, art. 22 (Aug. 5, 1990).

176. Victoria Nuland, *Press Statement on the ASEAN Declaration on Human Rights*, U.S. State Department (Nov. 20, 2012), https://2009-2017.state.gov/r/pa/prs/ps/2012/11/200915.htm [https://perma.cc/ZC2W-WNTC] ("While part of the ASEAN Declaration adopted November 18 tracks the [Universal Declaration of Human Rights], we are deeply concerned that many of the ASEAN Declaration's principles and articles could weaken and erode universal human rights and fundamental freedoms as contained in the UDHR.").

177. Aswad & Kaye, *supra* note 142.

178. Antoine Buyse, *Tacit Citing: The Scarcity of Judicial Dialogue between the Global and the Regional Human Rights Mechanisms in Freedom of Expression Cases*, *in* The United Nations and Freedom of Expression and Information: Critical Perspectives 443 (Tarlach McGonagle & Yvonne Donders eds., 2015) (describing these dynamics).

179. Int'l Law Comm'n, *Fragmentation of International Law: Difficulties Arising from the Diversification and Expansion of International Law*, ¶¶ 165-71, U.N. Doc. A/CN.4/L.682 (Apr. 13, 2006).

to some international norms than others because their domestic constitution incorporates some human rights treaties and not others, for example.[180] Overall, different authorities give priority to the norms that are, for various reasons, more relevant to their decisionmaking process.

This strategy is not available to corporate actors because they are not part of any treaty-based system. It is therefore unclear what rules, if any, should take priority in guiding their internal speech governance. The Board has so far applied U.N. treaties and regularly references the opinions and reports issued by U.N. bodies.[181] When Facebook and Twitter have referenced both U.N. and regional treaties, Aswad has called on them to be cognizant of the tensions between those instruments and to prioritize the protection that the U.N. affords to freedom of expression.[182]

IHRL advocates have looked for principles within international law to justify the preeminence of U.N. norms. International law, however, offers little help, as it contains only a few and not robust principles to solve conflicts of norms.[183] A report of the International Law Commission (hereinafter "ILC"), finalized by Martti Koskenniemi, addressed the fragmentation of international law.[184] The report's starting point is that "[w]hereas domestic law is organized in a strictly hierarchical way, with the constitution regulating the operation of the system at the highest level, there is no such formal constitution in international law and, consequently, no general order of precedence between international legal rules."[185]

The report by the ILC identifies legal techniques capable of resolving normative conflicts by putting rules in relationship with each other.[186] A few things are somewhat clear, such as the fact that the United Nations Charter takes priority over other treaties.[187] But the relationship between global and regional human rights treaties is not one of them. The report itself warns that looking for formal unity in a pluralistic and complex global society is "pointless."[188] Instead, fragmentation and normative conflicts reflect the differing preferences and projects of actors in a heterogeneous world. Identifying plausible ways to deal with these conflicts, however, may be valuable and even necessary for tribunals adjudicating conflicts.

---

180. Joana Harrington, *The Democratic Challenge of Incorporation: International Human Rights Treaties and National Constitutions*, 38 VICTORIA U. WELLINGTON L. REV. 217 (2007).

181. *See, e.g., Case decision 2022-003-IG-UA*, OVERSIGHT BD. (June 13, 2022), https://oversightboard.com/decision/IG-2PJ00L4T/ [https://perma.cc/CZD5-VMA4] (citing Communication 488/1992, Toonen v. Australia, Human Rights Committee, 1992; Resolution 32/2, Human Rights Council, 2016; UN Special Rapporteur on freedom of opinion and expression, reports: A/HRC/38/35 (2018) and A/74/486 (2019); UN High Commissioner for Human Rights, report: A/HRC/19/41 (2011)).

182. Aswad, *supra* note 23, at 44.

183. *See* Int'l Law Comm'n, *supra* note 179, ¶ 26.

184. *Id.*

185. *Id.*, ¶ 324.

186. Int'l Law Comm'n, *supra* note 179, ¶ 410.

187. U.N. Charter art. 103.

188. Int'l Law Comm'n, *supra* note 179, ¶ 16.

Most relevant to the problem of regional and universal treaties is the tool of lex specialis.[189] This technique prioritizes the most specific rule when two legal provisions are applicable and no clear hierarchical relationship exists between them. The closer connection to the particular context may more adequately reflect the interests and consent of the parties involved.[190] However, this technique did not make it into the Vienna Convention on the Law of Treaties from 1960, which identifies other methods for treaty interpretation.[191] Some authors dismiss this technique as irrelevant or impractical.[192] In any case, using lex specialis as a plausible tool to put global and regional rules in relation to each other is incompatible with the general rule that IHRL advocates propose (i.e., that U.N. norms always take priority).

Because international law has not traditionally dictated that U.N. norms should be preferred to regional norms, proponents of the IHRL project have focused on alternative interpretations of the relationship between human rights treaties. For example, Aswad's reply to Douek's objection about inconsistencies between regional and global treaties is that "[s]uch a concern inappropriately conflates IHRL with separate bodies of law embodied in regional human rights instruments."[193] Excluding regional human rights treaties from international law is highly uncommon. Describing a norm as "international" refers not to its global application, but to the source of legal authority upon which the norm exists.[194] That is, regional human rights treaties are part of international law "because they have been recognized as an international legal obligation through established international legal process."[195]

IHRL advocates also argue that "[r]egional human rights norms cannot, in any event, be invoked to justify departure from international human rights protections."[196] But this says nothing about the hierarchy of international norms or which rules companies should use to govern speech. This only says that a U.N. body would consider a state to be in violation of its international duties even if its conduct is permissible under another international norm. A regional tribunal could reach a similar conclusion if a state

189. *Id.*, ¶¶ 103-07.

190. *Id.*, ¶ 206; Hugo Grotious, *De Jure belli ac pacis. Libri Tres*, *in* The Classics of International Law 428 (James Brown Scott ed., 1925) ("Among agreements which are equal . . . that should be given preference which is most specific and approaches most nearly to the subject in hand, for special provisions are ordinarily more effective than those that are general.").

191. Vienna Convention on the Law of Treaties (May 23, 1969), arts. 31-33.

192. *See, e.g.,* Daniel P. O'Connell, International Law 253 (2nd ed. 1970).

193. Aswad, *supra* note 23, at 57.

194. Douglas Lee Donoho, International Human Rights Law 9 (2017).

195. *Id.*

196. SR Report 2019, *supra* note 132, ¶ 26. *But see* Kaye, *supra* note 46 (arguing that companies should draw guidance from 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 U.N. and regional human rights mechanisms).

adopted a rule that is compatible with a U.N. norm but incompatible with its duties under a regional treaty.

It is common for IHRL advocates to argue that the U.N. system acts as a floor or minimum of rights that companies must respect. Companies should only apply regional norms when these norms expand rights. Along those lines, Aswad argues that inconsistencies between international and regional mechanisms do not render the U.N. human rights regime incoherent: "[i]t simply means that the U.N. system provides more protections for speech than the regional systems."[197] Discussing the idea of referencing regional human rights treaties in the decisions of the Board, Co-Chair Catalina Botero stated that the Board could apply regional treaties in the future if the treaties expanded the level of protection of freedom of expression.[198]

Usually, the expansion of one right bears costs on other rights. For example, in regulating prior restraint on speech, the American Convention on Human Rights is more protective of speech than any other system.[199] But this protection of speech comes at the expense of less protection for other rights such as privacy or safety.[200] One can only see the Inter-American Convention as expanding rights if one assumes that the right that ought to be expanded is freedom of expression and not others.

In the context of the Board, this preference to expand freedom of expression over other rights has normative support. The Board's Charter refers only to "human rights norms that protect freedom of expression."[201] As a general rule for how companies should adopt international human rights, however, a normative basis is lacking. The principle is articulated in neutral terms ("apply the rule that offers higher protection") but it hides a clear normative orientation ("always choose the most free-speech protective rule").

Finally, IHRL advocates may find a basis for the preeminence of U.N. treaties in the UNGPs. The UNGPs only refer to the International Bill of Rights, which consists of the Universal Declaration of Human Rights, the ICCPR, and the International Covenant on Economic, Social, and Cultural Rights.[202] Therefore, Aswad argues that companies ought to look exclusively at global treaties. For example, she considers that Twitter's statement about

197. Aswad, *supra* note 136, at 642.

198. *Online Event: The Decisions of Facebook's Oversight Board – Implications for the Global South, particularly in Latin America*, THE DIALOGUE (May 18, 2021), min. 55:00, https://www.thedialogue.org/events/online-event-the-decisions-of-facebooks-oversight-board-implications-for-the-global-south-particularly-in-latin-america/ [https://perma.cc/92BB-QKSK].

199. American Convention on Human Rights, art. 13, Nov. 22, 1969, O.A.S.T.S. No. 36, 1144 U.N.T.S. 123 ("[t]he exercise of the right [to freedom of expression] shall not be subject to prior censorship but shall be subject to subsequent imposition of liability . . . [with the exception that] public entertainments may be subject by law to prior censorship for the sole purpose of regulating access to them for the moral protection of childhood and adolescence.").

200. Dvoskin, *supra* note 145.

201. *Oversight Board Charter*, art. 2(2), OVERSIGHT BD., https://oversightboard.com/governance/ [https://perma.cc/43ZW-2CMU].

202. Ruggie, *supra* note 84, at 19.

looking at U.S. law and the European Convention on Human Rights "departs from the UNGPs, which provide that companies should seek to align their operations with IHRL rather than domestic laws (like the U.S. Bill of Rights) or regional law (such as the European Human Rights Convention)."[203]

As argued above, this is a new interpretation of the UNGPs.[204] John Ruggie, the author of the UNGPs, believed companies should look at the International Bill of Rights and the ILO's Declaration on Fundamental Principles and Rights at Work as an authoritative list of recognized rights.[205] However, this did not mean that companies should follow the law stemming from U.N. treaties and their authoritative interpretation. In addition, discussing the relationship between global and regional norms, Ruggie concludes that "no homogenous hierarchical meta-system is realistically available within the international legal order to resolve the problem of incompatible provisions among different bodies of law."[206]

Looking at the efforts to build a coherent and uniform system of international rights makes the work toward objectivity visible. A clear relationship between global and regional norms creates the illusion that there is a system of norms readily available to constrain corporate power. However, asking companies to behave as if a coherent universal system of human rights existed gives corporations unprecedented powers. Corporations are put in charge of deciding which rights should be prioritized when international norms conflict, or how conflicts of norms between the European system of human rights (or any other regional system) and the U.N. should be adjudicated.

C. *Normative Indeterminacy as Technical Questions*

Scholars agree that IHRL has areas of significant indeterminacy.[207] Some scholars argue that technological innovation exacerbates this indeterminacy: the governance of speech on social media platforms poses novel questions that international law has not had the chance to address yet.[208] In addition, some believe that the speed and reach of the distribution of online expression make social media content exponentially more harmful.[209] Therefore, social media may require rules that give more weight to the values of dignity, safety, or equity. The question for the IHRL project is then, given the technical specificities of social media, should companies interpret IHRL dif-

203. Aswad, *supra* note 23, at 44; *see also* Aswad & Kaye, *supra* note 142, at 56.
204. *See supra* Section II.A.1.
205. Ruggie, *supra* note 84, at 19.
206. *Id.* at 65.
207. *See, e.g.,* Danielle Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1063 (2018); Douek, *supra* note 13, at 66.
208. *See, e.g.,* Douek, *supra* note 30, at 56, 72.
209. Danielle Citron, *Cyber Civil Rights*, 89 B.U.L. REV. 61, 63 (2010).

ferently than states and international bodies do when they restrict speech offline?

The Board has started to deal with this question. In some cases, it allowed Facebook to adopt rules that, in its own view, directly contradict what IHRL prescribes. Even in these cases, the Board justified these rules as an appropriate implementation of human rights in part because online speech is different from offline speech. A series of decisions evaluated Facebook's ban on a list of racial slurs[210] and Case Decision 2021-002-FB-UA reviewed Facebook's prohibition on content depicting blackface.[211]

The Board had to justify why, in its view, a state would violate international law if the state issued a ban on blackface or on racial slurs, but Facebook was meeting its human rights responsibilities when issuing those same bans. In the case about blackface, the Board hinted at four rationales: U.N. experts and other authorities determined that blackface creates objective harm;[212] Facebook has a human right responsibility to promote equality;[213] it is hard to evaluate the intent of the speaker on social media;[214] hate speech, even when it does not have an intent to discriminate or to incite violence, can create a discriminatory, harassing, and degrading environment.[215]

Consider the last two rationales. One could challenge them on factual terms. It might well be that racial slurs and people in blackface create a discriminatory environment regardless of whether the expression is online or off, and the intent of the speaker may always be hard to judge. It is also unclear that these types of expressions when they take place at one's university, in the workplace, or other close social environment are less harmful than when they occur on social media. But let us accept these distinctions between online and offline speech. Perhaps digital racial slurs and people in blackface create an even worse discriminatory environment because of their reach and quantity. What should we make of that difference?

The Board gave three different answers. First, the general ban on blackface that Facebook adopted is incompatible with international law.[216] In the Board's words, "international human rights law would not allow a state to impose a general prohibition on blackface through criminal or civil sanctions, except under the conditions foreseen in ICCPR Article 20, para. 2 and Article 19, para. 3." In the Board's view, the post discussed in this case

210. *Case decision 2021-011-FB-UA*, Oversight Bd. (Sept. 28, 2021), https://oversightboard.com/decision/FB-TYE2766G/ [https://perma.cc/UE9F-FHP7]; *Case decision 2020-003-FB-UA*, Oversight Bd. (Jan. 28, 2021), https://oversightboard.com/decision/FB-QBJDASCV/ [https://perma.cc/DC5H-STWU].

211. *See supra* Introduction (discussing this decision).

212. *Case decision 2021-002-FB-UA*, *supra* note 1, at 16.

213. *Id.*

214. *Id.*

215. *Id.*

216. *Id.* at 14.

"[w]ould fall within this category of protection from state restriction." Second, in this new medium, speech produces different consequences, and new answers are possible.[217] Third, the prohibition is compatible with international law. The Board concluded that "Facebook followed international guidance and met its human rights responsibilities in this case."[218]

Notice the tension between asserting that states would breach international law if they adopted this rule (that is, IHRL does prescribe what rules are acceptable) and that the new medium creates a situation in which new normative answers are possible (that is, IHRL does not prescribe what rules are acceptable).

The Board had two options to avoid that tension. First, given the open texture of IHRL, the Board could have argued that IHRL allows prohibitions on blackface, at least in the online context.[219] However, that would have meant committing to the view that states can also issue these regulations. What seems to animate the Board's reasoning is the belief that states and companies should actually govern speech differently. If that is the case, the Board also had a second option. It could have stated that IHRL was not the appropriate framework in this case because states and corporations are different. However, that would have meant losing the legitimizing force of IHRL.

Ultimately, the Board adopted a third choice: states and corporations ought to regulate speech differently and both options are compatible with IHRL. In order to preserve the claim to objectivity, the Board framed this third option as determined by technical facts.

The strategy here was to transform an unanswered question into one that can be answered with technical facts. The Board's move was the following: international law does not allow general bans on content depicting blackface, but online speech is more harmful and harder (perhaps impossible) to adjudicate on a case-by-case basis. The Board's conclusion could have been that IHRL provides no guidance to decide the case. Instead, the Board decided that these technical facts make Facebook's rule proportional and compatible with IHRL. Through these steps, the Board framed an open normative question as a question about how to translate international law principles in a new technical context, as if the technical context could determine the answer to the question. As a result, the Board did not appear to be creating new human rights standards, but rather appeared to be an objective translator.

Labeling new norms designed for platform governance as IHRL is a clear expression of the ideal of expert governance. Even if these new norms oppose IHRL norms, preserving the human rights label helps to maintain IHRL's

217. *Id.* at 16 (assessing the cumulative effects of speech on social media and the challenges of enforcement at scale).

218. *Id.* at 14.

219. Thanks to Evelyn Douek for raising this point in a personal conversation.

claims to legitimacy, universality, and democratic buy-in. This import of credentials from some initial principles (such as IHRL) to new normative outcomes is the essence of how expert governance works.

D. Local Preferences as Local Facts

Differences across geographical contexts bring into question the universality of human rights.[220] This tension between the universal character of human rights and local divergences is an old dispute in the field.[221] Scholars sometimes see the variations in normative solutions as a positive development.[222] Jack Goldsmith and Tim Wu value the different attitudes toward proper speech regulation as a reflection of differences in culture and taste.[223] More generally, the ILC finds that regional law better reflects the interests of the affected constituencies.[224]

The challenge for the IHRL project is how to account for the contradictions across cultures and societies while maintaining the claim that IHRL is shared across communities. A tool used to overcome this challenge is to reconstruct the variations across societies as problems of fair implementation. Indeed, the IHRL project often reduces the divergences between normative preferences to differences in the local facts. In turn, this move shifts the problem from the unavailability of universal normative solutions to a problem of lack of sufficient knowledge about the situation on the ground to ensure the correct application of global rules.[225] In order to address the implementation problem, the IHRL project emphasizes the importance of that local expertise. Local experts can make local contexts explainable and visible so that the global executor of rules can govern fairly.

This technique is not exclusive to the IHRL project. The broader project of creating a set of global rules often tries to incorporate different local preferences through expertise. Trusted partner programs are a prominent mechanism to incorporate local expertise in global content moderation.[226] Local NGOs assist in monitoring social media and applying companies' commu-

220. Siva Vaidhyanathan, Antisocial Media: How Facebook Disconnects Us and Undermines Democracy 27 (2018).

221. Kennedy, *supra* note 156.

222. *See, e.g.,* Dina 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); Malcom Evans, *The Future(s) of Regional Courts on Human Rights*, *in* Realizing Utopia 261 (Antonio Cassese ed., 2012).

223. Jack Goldsmith & Tim Wu, Who Controls the Internet? Illusions of a Borderless World 150 (2006).

224. Int'l Law Comm'n, *supra* note 179, ¶ 206.

225. *See* Stephen Humphreys, Theatre of the Rule of Law: Transnational Legal Intervention in Theory and Practice (2010) (exploring how rule of law promotion programs reframe political preferences as questions of expertise).

226. Naomi Appelman & Paddy Leerssen, *On "Trusted" Flaggers*, Yale-Wikimedia Initiative on Intermediaries & Information (July 12, 2022), https://law.yale.edu/sites/default/files/area/center/isp/documents/trustedflaggers_ispessayseries_jul2022.pdf [https://perma.cc/NWP4-UXE8]; Brenda Dvoskin, *Social Media Platforms and Civil Society in Latin America: A View from the Nonprofit Organizations* (Centro de Estudios en Tecnología y Sociedad, Working Paper No. 20, 2020).

nity guidelines. These programs provide local civil society organizations a privileged channel to report content that companies should take down or appeal decisions when content has been eliminated incorrectly.[227] The director of a Latin American organization says in an interview that Facebook often consults them about the meaning of a slang phrase or the details of an ongoing protest, but the opportunities to challenge the rules of the company, which they see as over-censorial, are practically nonexistent.[228]

The Board is particularly emphatic about the importance of local contexts. When deciding cases, it often cites expert reports that explain what rules the situation on the ground requires.[229] The interest in understanding the local situation adequately is a step worth celebrating.[230] However, like trusted partner programs, the Board is more invested in learning local facts rather than recognizing the variations in preferences about how to balance free expression and other values.

The Board usually looks at local specificities under the proportionality analysis of Article 19 of the ICCPR. This proportionality test is an interesting site to locate different possible balances between rights (in the case about Black Pete, between the right to freedom of expression and the right to non-discrimination). Indeed, whether a speech restriction is proportional to achieve a specific goal is fundamentally a matter of normative taste. A proportionality analysis demands answering questions such as how many false positives are tolerable to prevent certain harm.[231] For example, a ban on nudity can pursue the goal of preventing all cases of nonconsensual distribution of intimate images. The ban will also affect content that is uploaded consensually. A more nuanced rule could affect less content but would be less effective in countering all instances of nonconsensual uploads since errors will be unavoidable. Whether the most expansive ban is proportional to the goal will depend on how a society values the protection of nudity online or how essential a society considers the avoidance of nonconsensual distribution to be.

227. *See, e.g., EFHR welcomed into Trusted Partner Channel of Facebook*, European Foundation of Human Rights (Jan. 29, 2018), https://en.efhr.eu/2018/01/29/efhr-welcomed-trusted-partner-channel-facebook/ [https://perma.cc/8M2Y-A4AU].

228. Dvoskin, *supra* note 226, at 8-9.

229. *See, e.g., Case decision 2021-010-FB-UA,* Oversight Bd. (Sept. 27, 2021), https://oversightboard.com/decision/FB-E5M6QZGA/ [https://perma.cc/GCE6-75XA]; *Case decision 2020-006-FB-FBR*, Oversight Bd. (Jan. 28, 2021), https://oversightboard.com/decision/FB-XWJQBU9A/ [https://perma.cc/4K9L-K2GA].

230. Chinmayi Arun, *Rebalancing Regulation of Speech: Hyper-Local Content on Global Web-Based Platforms*, Medium (Mar. 28, 2018), https://medium.com/berkman-klein-center/rebalancing-regulation-of-speech-hyper-local-content-on-global-web-based-platforms-1-386d65d86e32 [https://perma.cc/XUD4-3K9W] (stressing the importance of local contexts for platform governance).

231. *See generally* Evelyn Douek, *Governing Online Speech: From "Posts-As-Trumps" to Proportionality and Probability*, 121 Colum. L. Rev. 759 (2021); New York Times Co. v. Sullivan, 376 U.S. 254 (1964) (arguing, although not in these terms, that actual malice was the appropriate standard of review because it avoids false positives, even if that means allowing some false negatives).

However, this is not the kind of local context that the Board considers necessary to consider. For example, in Case Decision 2020-006-FB-FBR, the Board analyzed Facebook's removal of false information about COVID-19.[232] The balance between free expression and public health in the context of the COVID pandemic is a good example of where people and communities hold different views. Platforms, accused of contributing to the spread of misinformation, have aggressively targeted public health misinformation at least partially in response to the preferences of some segments of the public.[233]

In the post at issue, a user had stated (incorrectly) that a cure for COVID-19 was available, that the drug was harmless, and complained that the French authorities were not making the drug available in France.[234] Facebook explained to the Board that, following the opinion of consulted experts, it decided to take down all content stating that a cure for COVID-19 is available because other users may believe it and may disregard precautionary health guidance or may self-medicate as a consequence.[235] The Board disagreed that this was the right balance between freedom of expression and public health. The user was questioning a governmental policy and calling for a change.[236] The Board reasoned that the protection of this kind of political discourse was fundamental.[237] In addition, the Board was not certain that the post could contribute to imminent harm.[238]

In this case, the Board highlighted that the drugs referenced in the post were not available in France without a prescription.[239] It added that, "the alleged cure has not been approved by the French authorities and thus it is unclear why those reading the post would be inclined to disregard health precautions for a cure they cannot access."[240] As in other decisions, the Board was very interested in looking at the offline context of a post, but this context was limited to facts that experts can explain and the Board can consider when making its own normative choice.

IV. Participating in the IHRL Project

At times, the IHRL project calls for a conversation to develop new rules.[241] As discussed above, one strong justification for the IHRL project is

232. *Case decision 2020-006-FB-FBR*, *supra* note 229.

233. *See, e.g.,* Guy Rosen, *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19*, Meta (Apr. 16, 2020), https://about.fb.com/news/2020/04/covid-19-misinfo-update/ [https://perma.cc/42GN-6N83]; *COVID-19 Medical Misinformation Policy*, Google (May 20, 2020), https://support.google.com/youtube/answer/9891785 [https://perma.cc/MTZ7-ATWJ].

234. *Case decision 2020-006-FB-FBR*, *supra* note 229, at 3-4.

235. *Id.* at 8.

236. *Id.*

237. *Id.* at 10.

238. *Id.* at 8.

239. *Id.*

240. *Id.* at 8-9.

241. Douek, *supra* note 30, at 47.

its alleged potential to facilitate the conversation among multiple stakeholders.[242] Evelyn Aswad calls for multi-stakeholder initiatives to refine some aspects of the IHRL framework.[243] David Kaye insists on the importance of including more actors in the conversation.[244]

These proposals for multi-stakeholder conversations or consultation mechanisms envision a project of objective governance different from the pursuit of exogenous standards discussed so far. These initiatives aim at building what Sheila Jasaoff calls a view from everywhere.[245] In this version, legitimacy depends on all views having adequate representation. Many authors theorize that legitimacy stems from procedures that ensure that all positions are heard. Ronald Dworkin argued that the legitimacy of laws depends on everyone's opportunity to manifest their opposition to them.[246] Owen Fiss contended that the purpose of the right to freedom of expression is genuine collective self-determination.[247] Therefore, free speech regulations ought to ensure that all viewpoints are heard.[248] In Jürgen Habermas's work, rules of participation and public reasoning create an ideal speech situation that makes the outcome of a deliberation worthy of respect by all parties, regardless of the substantive content of the result.[249]

Unlike the claims that IHRL is universal because it is the outcome of a participatory process that took place in the past (e.g., "states have ratified human rights treaties" or "IHRL is the outcome of a global consensus"), this version of the IHRL project calls for a future participatory process to draft new rules. It is unclear in what institutional setting that conversation may occur or how IHRL may facilitate it. On the one hand, the IHRL project may contribute substantively to these dialogues. It demands transparency and public justifications for corporate practices.[250] Transparency requirements would enable more informed conversations. On the other hand, the calls for transparency are strong enough not to gain much from the added support from the IHRL project. IHRL advocates also argue that it would provide a "common conceptual language" to be in conversation.[251] The test set out by Article 19 of the ICCPR could serve as a reasoning process to guide the conversation. However, the test itself, stripped from

242. *See supra* Section II.A.5.

243. Aswad, *supra* note 23, at 57.

244. SR Report 2018, *supra* note 9.

245. Jasanoff, *supra* note 53, at 315.

246. Ronald Dworkin, *Foreword, in* Extreme Speech and Democracy (Ivan Hare & James Weinstein, eds. 2009).

247. Owen Fiss, *Free Speech and Social Structure*, 71 Iowa L. Rev. 1405, 1411 (1986).

248. *Id.* at 1421 ("To assess the validity of the state intervention the reviewing court must ask, directly and unequivocally, whether the intervention in fact enriches rather than impoverishes public debate.").

249. Jürgen Habermas, Moral Consciousness and Communicative Action 43 (Christian Lenhardt & Shierry Weber Nicholsen trans., 1990) (1983).

250. Land, *supra* note 40, at 288; Douek, *supra* note 30, at 48.

251. Sander, *supra* note 30, at 967.

other requirements, is so thin that it is unclear how it would play a role in channeling debates.

IHRL may enable a wider conversation about content governance because it requires that corporations create new participatory mechanisms. IHRL advocates derive this conclusion from their new interpretation of the UNGPs, in particular the principles that refer to human rights due diligence.

Principle 17 of the UNGPs prescribes that "[i]n order to identify, prevent, mitigate and account for how they address their adverse human rights impacts, business enterprises should carry out human rights due diligence."[252] Human rights due diligence should "[i]nvolve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation."[253]

The Board interprets these principles to mean that Facebook is expected to consult with civil society and other stakeholders when developing new content moderation rules or practices.[254] Initiatives to achieve legitimacy through these procedures exist everywhere. All major companies now have different forms of engaging external stakeholders in their internal rule-making process.[255]

The Board has shown an interest in mechanisms for civil society organizations to participate. The Board provides opportunities for the public to submit comments in all cases in a similar fashion to amici curiae.[256] In some instances, it has established rudimentary forms of dialogue with civil society organizations, directly citing reports or addressing concerns expressed by

252. U.N. Special Representative of the Secretary-General, *supra* note 68, Principle 17.

253. *Id.*

254. *See, e.g., Case decision 2021-011-FB-UA*, *supra* note 210, at 10.

255. *See, e.g.,* Vanessa Pappas, *Introducing TikTok Content Advisory Council*, TikTok (Mar. 18, 2020), https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council [https://perma.cc/DL34-C3M7]; Arjun Narayan Bettadopur, *Introducing the TikTok Asia Pacific Safety Advisory Council*, TikTok (Sept. 22, 2020), https://newsroom.tiktok.com/en-sg/tiktok-apac-safety-advisory-council [https://perma.cc/5DMU-YGNM]; Julie de Bailliencourt, *Meet TikTok's European Safety Advisory Council*, TikTok (Mar. 1, 2021), https://newsroom.tiktok.com/en-gb/tiktok-european-safety-advisory-council [https://perma.cc/83VD-NNWU]; *Introducing the Twitch Safety Advisory Council*, Twitch (May 14, 2020), https://blog.twitch.tv/en/2020/05/14/introducing-the-twitch-safety-advisory-council/ [https://perma.cc/YS3G-YCAK]; Nick Pickles, *Strengthening our Trust and Safety Council*, Twitter (Dec. 13, 2019), https://blog.twitter.com/en_us/topics/company/2019/strengthening-our-trust-and-safety-council.html [https://perma.cc/ZB4M-T759]; Patricia Carter, *Announcing the Twitter Trust & Safety Council*, Twitter (Feb. 9, 2016), https://blog.twitter.com/en_us/a/2016/announcing-the-twitter-trust-safety-council.html [https://perma.cc/UW3E-WSSS]; *Stakeholder Engagement*, Meta, https://www.facebook.com/communitystandards/stakeholder_engagement [https://perma.cc/KL63-KP64].

256. *See, e.g., Case decision 2020-006-FB-FBR*, *supra* note 229, at 15 (acknowledging that the board's recommendations to Facebook drew on public comments the board received).

civil society.[257] The Board has valued Facebook's efforts to engage civil society.[258]

As discussed above, the UNGPs refer to companies' responsibility to respect human rights that international law recognizes.[259] Accordingly, when the UNGPs discuss human rights due diligence, they envision that companies will make an effort to audit the impact of their operations on human rights.[260] For example, under this framework, Facebook is expected to understand how its platforms are used in different countries. If it is the case, as it has been in Myanmar, that the platform is a key tool in organizing crimes against humanity, Facebook should be aware of this and take active steps to prevent or remedy its involvement in human rights abuses.[261]

The Board's interpretation goes well beyond that responsibility. When Facebook engages civil society around the world to gather feedback on how to amend its community guidelines, it proposes different potential rules and hears advantages or concerns about each one of them.[262] The kind of stakeholder engagement that companies carry out and that the Board promotes does not have the sole or even central purpose of identifying potential human rights infringements. Rather, companies engage civil society to incorporate their preferences among multiple human-rights-respecting policy options.[263] The assumption is that all discussed alternatives respect human rights.

Even if divergent from the text of the UNGPs, the Board's interpretation can be very useful. The main stated goal of the IHRL project is to align private speech regulation with the interests of the people. This reading of the human rights due diligence responsibility may strengthen companies' consultation with the public during the policy development process. Along those lines, in Case Decision 2020-006-FB-FBR concerning a post containing inaccurate information about COVID-19 treatments, the Board recom-

257. *See, e.g., Case decision 2021-011-FB-UA*, *supra* note 210, at 11 (citing a report elaborated by the NGO Media Monitoring Africa).

258. *See, e.g., Case decision 2021-002-FB-UA*, *supra* note 1, at 11.

259. *See supra* Section II.A.1.

260. *See* Ruggie, *supra* note 84, at 99.

261. *See id*. at 100 (describing the responsibilities of multinational corporations operating in countries where they are at risk of becoming complicit in egregious human rights abuses).

262. *See generally* Matthias Kettemann & Wolfgang Schulz, *Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study* (Hans-Bredow-Institut Working Paper, 2020) (describing the steps in Facebook's stakeholder engagement process).

263. *See, e.g.*, *Standing Against Hate*, META (Mar. 27, 2019), https://about.fb.com/news/2019/03/standing-against-hate/ [https://perma.cc/BRV3-D9A7]; Monika Bickert, *Enforcing Against Manipulated Media*, META (Jan. 6, 2020), https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/ [https://perma.cc/C2YX-PB7U]; Vanessa Pappas, *Combating Misinformation and Election Interference on TikTok*, TIKTOK (Aug. 5, 2020), https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok [https://perma.cc/6XJH-VDER]; Twitter Safety, *Updating Our Rules Against Hateful Conduct*, TWITTER (July 9, 2019), https://blog.twitter.com/en_us/topics/company/2019/hatefulconductupdate.html [https://perma.cc/PE5X-WD4L]; Vijaya Gadde & Del Harvey, *Creating New Policies Together*, TWITTER (Sept. 25, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Creating-new-policies-together.html [https://perma.cc/7EAB-DX6B].

mended that "Facebook should conduct a human rights impact assessment with relevant stakeholders as part of its process of rule modification" per Principles 18 and 19 of the UNGPs.[264] In Case Decision 2021-006-IG-UA concerning a post discussing the solitary confinement of Abdullah Öcalan, the Board recommended that Facebook "ensure meaningful stakeholder engagement" to review its policy on dangerous individuals and organizations, including through a public call for input.[265] Notice that in this case, the Board did not refer to any specific IHRL obligation. Perhaps the Board can make this recommendation in a similarly convincing fashion regardless of whether the Board refers to the UNGPs.

In the future, the Board's interest in stakeholder engagement may diverge further from the UNGPs' meaning and evolve into more precise standards for stakeholder consultation. For example, in Case Decision 2021-011-FB-UA, regarding a racial slur in South Africa, the Board appreciated Facebook's consultation with external stakeholders to draft an exception to its hate speech policy for insults when used self-referentially. The Board also pointed out that the external stakeholders represented diverse geographical regions.[266]

In some cases, the Board hinted timidly that consultation with civil society organizations could even lead the Board to accept rules that diverge from interpretations made by the Human Rights Committee or U.N. Special Rapporteurs. In Case Decision 2020-007-FB-FBR, regarding an alleged veiled threat, the minority was willing to defer to Facebook's decision because Facebook had consulted with regional and linguistic experts and had worked with a local partner to identify and adjudicate the content. In Case Decision 2021-002-FB-UA, regarding Facebook's ban on content depicting blackface, the Board took into account that the rule was the outcome of a process that "involved extensive research and engagement with more than 60 stakeholders, including experts in a variety of fields, civil society groups, and groups affected by discrimination and harmful stereotypes."[267] The majority considered this stakeholder consultation to be in line with "international standards for on-going human rights due diligence" and cited Principles 17(c) and 18(b) of the UNGPs.[268]

Despite the Board's interest in stakeholder engagement initiatives, the view from nowhere has, so far, prevailed in the Board's deployment of the IHRL project. In the case regarding Facebook's general prohibition on content depicting people in blackface, the Board appreciated the process of engaging stakeholders.[269] However, when justifying its decision to uphold the

264. *Case decision 2020-006-FB-FBR*, *supra* note 229, at 13.

265. *Case decision 2021-006-IG-UA*, Oversight Bd. (July 8, 2021), https://oversightboard.com/decision/IG-I9DP23IB/ [https://perma.cc/J2SD-6GZP].

266. *Case decision 2021-011-FB-UA*, *supra* note 210, at 7.

267. *Case decision 2021-002-FB-UA*, *supra* note 1, at 11.

268. *Id.*

269. *Id.* at 11.

ban, the Board did not reference this process. Instead, the Board decided that the divergence was justified because many experts had found that images depicting blackface are discriminatory and harmful. Among other reasons discussed above, the Board relied on the fact that "[n]umerous human rights mechanisms have found the portrayal of Zwarte Piet to be a harmful stereotype."[270] But not any harmful stereotype would satisfy the Board. The Board cited many reports and concluded that these expert findings provided "sufficient evidence of *objective* harm to individual's rights to distinguish this rule from one that seeks to insulate people from *subjective* offense."[271] This passage defines a boundary between objective knowledge and subjective feelings. Here, objectivity comes not from the stakeholder engagement process but from the human-rights experts' findings.

Another example in the same direction comes from Case Decision 2021-011-FB-UA concerning the use of a racial slur in South Africa.[272] The Board analyzed Facebook's decision to delete a post that used a racial slur in the context of discussing wealth and racial dynamics in South Africa. The Board had already decided other cases dealing with bans of specific terms, always finding that even though prohibiting the use of particular words would breach IHRL if adopted by a state, Facebook's ban was compatible with IHRL.[273]

This case offers an additional glimpse of how the Board divides tasks among U.N. experts and civil society organizations. On the one hand, the Board stated that banning racial slurs is incompatible with IHRL. Still, Facebook can adopt that ban because the "[U.N.] Special Rapporteur indicates that entities engaged in content moderation like Facebook can regulate such speech."[274] On the other hand, the Board stated that Facebook should consult affected groups and human rights experts, as it did in this case, to review its policies and update the list of banned slurs.[275] The consultation in this case contributed to Facebook's understanding of the meaning of the term in the context where it is used.[276] In this sense, engagement with local civil society is used mainly as an implementation mechanism to ensure the correct understanding of racial slurs, while the rule-making process is kept in the hands of a small group of experts.

270. *Id.* at 14.
271. *Id.* at 16 (emphasis added).
272. *Case decision 2021-011-FB-UA*, *supra* note 210.
273. *Case decision 2020-003-FB-UA*, *supra* note 210; *Case decision 2020-007-FB-FBR*, Oversight Bd. (Feb. 12, 2021), https://oversightboard.com/decision/FB-R9K87402/ [https://perma.cc/9VT9-XDYP].
274. *Case decision 2021-011-FB-UA*, *supra* note 210.
275. *Id.*
276. *Id.*

V. The Futures of the IHRL Project

Overall, what is problematic, if anything, with expert governance? This section asks if the IHRL project might indeed help us bridge the democratic deficits that David Kaye identifies. Because the IHRL project is still incipient, the answer might depend on how the IHRL project evolves. Therefore, this section answers that question by imagining what I believe are the most auspicious futures for the IHRL project. First, it might be that human rights experts persuade many people that their decisions are good. Second, it might be that the participatory aspects of the project open up space for more actors and flatten the power relationships among governance actors. Finally, I try to imagine what disentangling content moderation from IHRL could look like. I suggest that the execution of the IHRL project has so far undermined its own stated goals, and that this disentanglement might offer a more promising path.

A. A Communal Viewpoint Developed from the Top

IHRL skeptics, myself included, have raised the concern that the main effect of the IHRL project may be to legitimize corporate decisions without imposing severe constraints on their power.[277] From the perspective of the Board or the advocates of the IHRL project, having human rights experts with perceived legitimacy to govern speech might be a good result. If using the IHRL framework brings about legitimacy either to the Board or to companies framing their policies in human rights language, this outcome could be considered a victory.

Even from the perspective of the public at large, it is not obvious why this outcome of increased legitimacy would be problematic. If the public perceives experts as legitimate governors, it might be because the public finds the experts' actions and decisions convincing. Invoking IHRL will likely help the experts make their decisions more compelling. It is possible that, over time, experts persuade a large number of people that the solutions that they offer are reasonable. In that case, the Board's legitimacy might come from the public's approval and agreement with the Board's decisions. In the long term, it could be that the Board's decisions reflect a communal viewpoint developed from the top.[278]

Some of the risks of making IHRL the relevant and legitimate language of online speech governance are analogous to the risks that the human rights movement faces more generally. If IHRL is perceived as the yardstick for online speech governance, there might be less intellectual energy to concep-

277. *See, e.g.,* Douek, *supra* note 30, at 63; Dvoskin, *supra* note 145.

278. *See* Sheila Jasanoff, Designs in Nature: Science and Democracy in Europe and the United States 152-55 (2005) (explaining and illustrating the idea of a communal viewpoint developed from the top).

tualize other normative possibilities.[279] Relatedly, Barrie Sander highlights that the inevitable risk of the IHRL project is that it may legitimize minor improvements "at the expense of undertaking more structural" changes.[280] Looking at transparency and due process through the lens of human rights carries the risk of focusing too strongly on formal procedures while ignoring the power differential among users and civil society to make use of these mechanisms.[281]

More important is IHRL's ambition to reflect global preferences. Ultimately, the faith or lack thereof in the IHRL project might reside in whether one believes that speech regulation should aim to reflect a broad global consensus—a communal viewpoint—or that speech regulation is mainly a permanent site of disagreement. Daniel Walters, reflecting on the relationship between administrative agencies and democracy, criticizes agencies for "trying to convince participants that the agency's proposal is good for everyone."[282] Instead, he argues, agencies ought to "forthrightly acknowledge that the proposal may *not* be good for everyone."[283] Likewise, performing as executors of some objective notion of the will of the global society denies the distributive effects of experts' decisions and the fact that some viewpoints will be necessarily disadvantaged. In turn, that claim to universality conceals the need for stronger mechanisms to share power among diverse constituencies.

B. IHRL as a Participatory Project

The promise of a broader conversation with the power to shape content moderation facilitated by IHRL has yet to be realized. I am skeptical that a shared language, common framework, or proportionality test can meaningfully impact the power distribution among the actors that can or do participate in online speech governance. However, existing institutions may leverage the IHRL project to create or strengthen institutional initiatives to increase the number of opportunities for participation. Additionally, it could be that engaging in IHRL discourse helps these institutions build their credibility, which they might in turn use to pursue projects unrelated to IHRL.

The principal executor of the IHRL project has been the Board. So far, it has found experts' reports, especially from U.N. bodies, to be the most promising source to build its own legitimacy. As the Board matures, it

279. *See* Kennedy, *supra* note 156, at 108; Samuel Moyn, Humane: How the United States Abandoned Peace and Reinvented War (2021) (arguing that international humanitarian law diverts from pursuing more radical projects).

280. Sander, *supra* note 30, at 1005.

281. *See* Ganesh Sitaraman, *The Puzzling Absence of Economic Power in Constitutional Theory*, 101 Cornell L. Rev. 1445, 1500 (2016); Nico Krisch, *The Pluralism of Global Administrative Law*, 17 Eur. J. Int'l. L. 247, 276 (2007).

282. Walters, *supra* note 57, at 56.

283. *Id.* (emphasis in the original).

might be willing to lend some power to stakeholder engagement initiatives. The Board could show more deference to Facebook's rules when these rules are the product of consultations with civil society. The Board should develop standards for such consultations in terms of how open they need to be, whose participation should be ensured, how transparent the process needs to be, and how the company needs to consider the received input.[284] The Board could ask for a report of the disagreements identified in the stakeholder engagement process, and which stakeholders were favored by the outcome of the policy-drafting process. The Board may ground these standards in its innovative interpretation of the UNGPs on human rights due diligence discussed above[285] or use the legitimacy that human rights discourse brings to undertake these participatory projects.

The Inter-American Commission on Human Rights has recently launched a multi-stakeholder dialogue on making content moderation policies "compatible" with IHRL standards.[286] This initiative contemplates multiple opportunities to seek input from experts, the general public, and other regional and international endeavors. It could be a significant effort to set authoritative standards for content moderation developed outside of companies, borrowing legitimacy from participatory procedures and using the IHRL framework to strengthen the authority and persuasiveness of the substantive outcomes.

This broadening in institutional opportunities is a positive development. On the one hand, the Commission's initiative will not necessarily address the background conditions that explain why civil society across geographical areas and ideological viewpoints has different capacities, resources, and leverage to influence online speech governance.[287] As a result, it might create more formal participatory mechanisms but it might also be insufficient to ensure that those mechanisms are used equitably.[288] On the other hand, the initiative can turn up the volume of Latin American voices, which have so far been neglected in global debates.

Ultimately, these participatory initiatives will have to be evaluated based on their own merits, not on how closely they follow IHRL guidance. Their main weakness is that their processes are designed to find convergence around online speech norms. The risk is confining participation within narrow bounds as a ritualized phase of policy development or a one-time dialogue.

---

284. *See* Dvoskin, *supra* note 226.

285. *See supra* Section IV.B.

286. *Americas Dialogue on Freedom of Expression Online*, AMERICAS DIALOGUE, https://www.americasdialogue.org/en/americas-dialogue/ [https://perma.cc/F68V-E88R].

287. *See* Dvoskin, *supra* note 226, at 10.

288. *See* K. Sabeel Rahman, *Policymaking as Power-Building*, 27 S. CAL. INTERDISC. L.J. 315, 353 (2018) (arguing that participatory mechanisms need actors willing and able to make full use of them).

C. Disentangling Content Moderation from IHRL

The IHRL project provides a framework for experts to present their decisions as objective to the public and, in turn, gain the public's acceptance. This is a tempting offer. But resisting the comfort that legal language provides and coming to terms with one's power can lead to decisions that are more transparent, less likely to be framed as universal answers, and easier to contest.

If newly empowered experts are too concerned with proving their objectivity, they may miss a valuable opportunity for plural and experimental governance. Indeed, the most interesting recommendations the Board has made to date were not grounded in IHRL.[289] Moving away from the IHRL project may prompt experts to provide better justifications for their decisions, involve more actors, and make space for more meaningful inclusion of normative disagreements. Imagining how the Board could have reached its decision in the case about Black Pete illustrates these points.

First, IHRL is highly praised for offering a framework to justify decisions to the public.[290] Likewise, one of the core purposes of the Board is to deliver transparent and well-justified decisions.[291] But in the Black Pete case, the Board's adoption of the IHRL project actually harmed the coherence and clarity of its decision.

Throughout the Article, I have discussed how the Board dealt with technical facts,[292] U.N. experts,[293] and stakeholder engagement in the Black Pete case.[294] An additional point to consider is how the Board assigns corporate actors a fundamental role in the protection of human rights that, in the Board's view, is a role barred to states. In the Black Pete case, the Board concluded that "international human rights law would not allow a state to impose a general prohibition on blackface"[295] and at the same time asserted that "Facebook followed international guidance and met its human rights responsibilities in this case."[296]

---

289. *See, e.g., Case decision 2021-009-FB-UA*, Oversight Bd. (Sept. 14, 2021), https://www.oversightboard.com/decision/FB-P93JPX02/ [https://perma.cc/NR3T-WM65] (recommending Facebook ". . . conduct a thorough examination to determine whether Facebook's content moderation in Arabic and Hebrew, including its use of automation, have been applied without bias."); *Case decision 2020-004-IG-UA*, Oversight Bd. (June 28, 2021), https://oversightboard.com/decision/IG-7THR3SI1/ [https://perma.cc/PNY8-B7HX] (recommending Facebook "[i]mplement an internal audit procedure to continuously analyze a statistically representative sample of automated content removal decisions to reverse and learn from enforcement mistakes."); *Case decision 2021-006-IG-UA*, *supra* note 265 (recommending Facebook open a public call for inputs on how to update its Dangerous Individuals and Organizations policy).

290. Sander, *supra* note 30, at 967.

291. Feldman, *supra* note 8, at 102.

292. *See supra* Section III.C.

293. *See supra* Section IV.

294. *Id.*

295. *Case decision 2021-002-FB-UA*, *supra* note 1.

296. *Id.*

The reasons why the Board thought that this was an appropriate rule for Facebook to adopt were, again, that blackface produces objective harm, that the real intent of the speaker on social media is hard to determine, that this kind of speech can create a discriminatory, harassing, and degrading environment, and that Facebook should promote equality.[297] It is hard to understand why Facebook would be expected to prevent the harms caused by this kind of content under its human rights responsibility but a state would not be allowed to take analogous steps under the same legal framework.

All of the circumstances that justified Facebook's decision also apply to state regulation. Instances of discriminatory speech, even if not intended to cause harm, are pervasive in many contexts and create discriminatory environments.[298] If experts' reports find that people in blackface create "objective" harm, then the harm is obviously independent of whether Facebook or governments are the ones to address it. States have a greater responsibility than Facebook to promote equality. Ultimately, the need to justify a rule that diverges from IHRL as an adequate application of that framework makes the decision more confusing, not more transparent.

Second, being transparent about what is driving the Board's decisions would enable more effective participation by the public. In the comments received by the Board in this case, Professor Sejal Parmar from the University of Sheffield was the only commenter who provided an analysis of the decision under IHRL, and particularly Articles 19 and 20 of the ICCPR.[299] Her comment cites some of the same reports the Board referenced, although she highlights that the reports do not call for a total ban on portraits of Black Pete. Ultimately, the Board decided to diverge from IHRL. In this case, telling the public that IHRL will determine the decision does not enhance the public conversation, as the project promises, but rather undermines it. Indeed, it becomes less clear to the public what reasons the Board will consider to be weighty when making a decision.

The Board instead could have explained why Facebook's rule aligned with the Board's normative approach. For example, the most important factor driving the decision could have been the support that the prohibition had among consulted stakeholders. If the Board acknowledges that this factor is driving its decision, then the Board can openly set requirements that a consultation process needs to meet in order to persuade the Board to give deference to the consultation's findings.

Third, if the Board is successful in promoting IHRL as the correct moral yardstick for content moderation, other companies might believe they need

297. *Id.*

298. *See* Chester Pierce, *Psychiatric Problems of the Black Minority*, *in* AMERICAN HANDBOOK OF PSYCHIATRY 512 (Silvano Arieti ed., 1974) (coining the term "microaggressions" and studying their impact on mental health).

299. *Public Comment Appendix for Case decision 2021-002-FB-UA*, OVERSIGHT BD. (Apr. 13, 2021), https://www.oversightboard.com/sr/decision/2021/002/public-comments [https://perma.cc/3GB5-BD37].

to converge with the Board in their decisions. When the Board says that Facebook is meeting its human rights responsibilities by adopting a general ban on blackface, the Board makes it harder for other companies to adopt a different rule. More fundamentally, the Board's proposal for how to assign roles in online speech governance greatly increases the power of corporate actors (including the Board's own power). In the Board's interpretation, a great objective harm exists to social equality that only companies (and the Board) should remedy. Framing its decisions as the product of its members' views would help the Board avoid this kind of general and universalizing stance of what online speech governance should look like.

Finally, abandoning the ideal of objectivity would help the Board be more explicit about its political commitments. As Walters says about administrative agencies, there is a lot to gain from the Board admitting that its decisions reflect some people's preferences, not everyone's.[300] Instead of putting an end to the debate, making normative preferences explicit calls for instances for those with different preferences to contest the decision in the future. In expert governance, the quest for objectivity is an obstacle for meaningful participation. Ultimately, putting politics front and center may more effectively facilitate the public conversation, transparency, and public justifications that the IHRL project envisions.

Conclusion

The IHRL project for content moderation promises to rein in corporate power on behalf of the public interest. It rests on the idea that IHRL can function as an objective synthesis of the global public interest. The main goal of this Article has been to examine the tools that experts have developed to build that claim to objectivity.

300. Walters, *supra* note 57, at 56.