NEUTRAL GOVERNANCE

Brenda Dvoskin Dannecker

This paper challenges the idea that international human rights law can be an effective tool to realign the governance of online speech with the public interest. It argues that proponents of this framework imagine an apolitical neutral system of governance. This neutral system would be responsive to previous global commitments reflected in international law and open spaces for all viewpoints to influence governance more equally. The paper focuses on the work that goes into building a neutral governance framework. It warns that the project may legitimize experts' power at the expense of pluralistic politics.

Introduction 1

I. Why IHRL? 2

A. Because platforms are like states, but they are not like states 2
B. Because IHRL is consented to by all states, but state consent does not matter 4
C. Because IHRL constrains corporate power, but its indeterminacy is a feature 6
D. Because IHRL is a shared language and because IHRL is like the First Amendment 7

II. The Ideal of Neutrality 8

A. Self-evidently good 9
B. Consent 10
C. U.N. Guiding Principles 11
D. A problem of implementation 11

III. Neutrality and Participation 13

A. Top-down participation: between Dworkin and Habermas 14
B. Bottom-up participation: IHRL disrupting neutrality 17

Conclusion 18

INTRODUCTION

Mainstream discourse acknowledges that social media companies have too much power over online speech and need to be subjected to external forms of control that respond to the public interest. International human rights law (IHRL) is an intuitively attractive framework to rein in that corporate power: it is global like these platforms and offers well-respected standards to guide content moderation in these quasi-public spaces.[1] The project has gained more attention and traction since David Kaye proposed it in 2018 during his tenure as U.N. Special Rapporteur. In a nutshell, he recommends that large social media platforms adopt IHRL as their own internal rules to moderate content.[2] Throughout the paper, I refer to this proposal as the IHRL project. This year,

[1] Rikke Frank Jørgensen, *What Platforms Mean When They Talk About Human Rights*, 9 POLICY & INTERNET 219 (2017), https://doi.org/10.1002/poi3.152 (showing how companies present themselves as public spaces).

[2] 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*

with the launch of the Facebook Oversight Board, which uses IHRL to articulate its decisions, the project has taken off and began to shape content moderation.

This paper unpacks the intuition that IHRL is an appropriate framework to guide content moderation at scale. It looks at how Kaye and other advocates justify why IHRL is a suitable framework. The way that proponents discuss the virtues of IHRL as a framework for content moderation is frequently internally inconsistent. The first section looks at these conflicting claims about the status of IHRL as a global set of standards that can constrain corporate power on behalf of the public interest.

My aim is not to show which statements are accurate or inaccurate, solve these contradictions, nor coherently reconstruct these statements. Instead, I believe there is much to learn from these internal tensions. Mainly, I am interested here in seeing the ideals that underlie the different pieces of these justifications. In the second section, I argue that the various fragments of the rationales and even some of their criticisms share an ideal of neutral governance. By neutral governance, I mean a system of governance with no losers: an apolitical system of objective ruling where the correct conception of good prevails.

The third section complicates the ideal of neutral governance that I present in the second section. The IHRL project does not only import exogenous standards of what is good, it also proposes some initiatives for participation, especially participation of civil society organizations. Participation can build neutrality in governance through a different mechanism: it can promise there will be no losers because the substantive outcome will encompass all viewpoints. The first part of this section looks at how these two mechanisms of neutral governance compete within the IHRL project. In the second part of this section, I argue that new actors may be appropriating IHRL as a new language to influence governance. I conclude that although we can see some hints in this direction, IHRL's major work so far has been to legitimize expert decision-making power and shield it from participatory politics.

I. WHY IHRL?

Often, IHRL is invoked as the proper benchmark to evaluate content moderation without explaining why that would be appropriate. People sometimes assume IHRL to be self-evidently good. However, at other times proponents of the IHRL framework justify why it would be the right set of rules for platforms to decide how to moderate content. These reasons are, as I show in this section, internally inconsistent.

A. Because platforms are like states, but they are not like states

When scholars try to describe what is so upsetting about how companies moderate content, they often claim that companies are carrying out state functions without the safeguards that rein in government power. 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

and Expression, No. A/HCR/38/35 (Jun. 2018) [hereinafter SR *Report 2018*].

> active social media users…. In a reversal of the historic roles, private corporations have at times become the de facto regulators of government speech.[3]

It is common for scholars to 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."[4] Julie Cohen says, "Dominant platforms' role in the international legal order increasingly resembles that of sovereign states."[5] Daphne Keller agrees: "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."[6] Describing platforms as states has profound normative implications. Because we imagine that platforms do what states do, it could follow that the problem is that platforms act like states without subjecting themselves to the legal frameworks that apply to states.

Platforms, however, are not states. Accordingly, scholars propose adaptations to the framework. However, these adaptations risk leaving too little of IHRL standing. This manifests in three ways in the IHRL project. Article 19 of the International Covenant on Civil and Political Rights is the epicenter of the IHRL project. It sets out a test to evaluate restrictions on freedom of expression. The test includes three requirements: rules must be prescribed by law (legality), must have a legitimate aim (legitimacy), and must be necessary for that aim (necessity). Proponents of the IHRL project acknowledge that the first two requirements need to be "translated" because platforms are not states.

Proponents of the IHRL project struggle to define what would be legitimate aims for content moderation. The aims that Article 19 defines as legitimate are over- and under-inclusive for platforms. On the one hand, Susan Benesch argues that firms are not well positioned to determine, for example, national security or public health goals.[7] On the other hand, Evelyn Aswad asks if it would be legitimate for companies to pursue the aim of maximizing profit.[8] 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, this specific discussion about commercial aims may not be that relevant. Companies will usually find it very easy to come up with a public-interest rationale to restrict speech without invoking business reasons. What is problematic is that IHRL, as adapted for platforms, may accept any proposed aim as legitimate. In fact, in all decisions to date, the Facebook Oversight Board has found that Facebook's community standards have a legitimate aim, including a case reviewing Facebook's super-expansive ban on adult nudity.

---

[3] Richard Ashby Wilson & Molly Land, *Hate Speech on Social Media: Towards a Context-Specific Content Moderation Policy*, 52 CONN. L. REV. 1, 5 (2020) (footnotes omitted).

[4] Nadine Strossen, *United Nations Free Speech Standards as the Global Benchmark for Online Platforms' Hate Speech Policies*, 29 MICHIGAN STATE UNIV. COLL. L. INT'L L. REV. 307, 324-325 (2021).

[5] Julie Cohen, *Law for the Platform Economy*, 51 U.C. DAVIS L. REV. 133, 199 (2017).

[6] Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech* 2-3 (Hoover Inst. Aegis Paper Series, Paper No. 1902, 2019).

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

[8] Evelyn Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26 (2018).

The requirement of legality has also been silently emptied. The condition demands that restrictions to freedom of expression are provided by *law*. That is, only bodies authorized to make law may impose such limits—typically parliaments or judicial courts. That is impossible within the IHRL project because companies are not states, and their structures do not incorporate bodies with those legitimate credentials. Accordingly, the requirement has been narrowed down to its substantive aspect, which refers to the clarity and precision of legitimate restrictions. A requirement that refers to the authoritative origins of rules is narrowed down to a requirement about their transparency.

Overall, scholars imagine social media companies as states, making the legal frameworks designed for states readily applicable. On closer examination, these companies do not have democratic representative bodies nor the legitimacy to make determinations about goals regarding public health, national security, or the appropriate delimitation between free speech and other interests. So even if companies do what states ought to do, they do not have the institutional structures or features that are critical pieces to operate the IHRL machinery.

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

Outside of the IHRL debate, everyone seems to be well aware that no universal, or even local, agreement on how to regulate speech exists. However, when justifying the IHRL project, advocates remind us that it is the only globally-consented framework. In his 2018 report, Kaye stated, "The 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."[9] IHRL promises a universal basis to moderate content.[10] As Aswad says, "Companies 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."[11]

IHRL's claim to universality comes from state consent, particularly the vast adoption of U.N. treaties.[12] As Nadine Strossen puts it, "almost every single country in the world is a party to the ICCPR [International Covenant on Civil and Political Rights] and the ICERD [International Convention on the Elimination of All Forms of Racial Discrimination]. Therefore, by moderating content in accordance with those U.N. treaties... Platforms would be honoring each country's international legal obligations."[13] The United Nations Guiding Principles on Business and Human Rights (UNGPs) are the IHRL project's other leg. In 2011, the Human Rights Council (HRC) endorsed this document that sets the expectations that all companies will respect human rights. As with the ICCPR, proponents of the IHRL project invoke their wide acceptance as one of the main

[9] *SR Report 2018*, *supra* note 2, ¶ 41 (internal footnotes omitted).

[10] *Id.*, ¶ 42.

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

[12] Aswad, supra *note 8*, at 35.

[13] Strossen, *supra* note 4, at 329.

reasons to acknowledge their legitimacy. Kaye says they establish "global standards."[14] Aswad describes them as reflecting "global expectations for companies to respect international human rights in their business operations."[15] Strossen and Aswad emphasize that the HRC adopted the UNGPs *unanimously*.[16]

From these statements, one could assume that states would be a good source for identifying a community's values or agreements. However, IHRL advocates also see state institutions as terrible sources for ascertaining the public interest—regardless of the state's democratic credentials. Indeed, IHRL flagrantly disregards states' lack of consent to particular rules. After explaining that the relevant standards that reflect global values can be found in human rights law, Kaye immediately adds that "Companies should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States."[17]

States are regarded with deep suspicion, and IHRL-making institutions are called to limit the opportunities for states to interpret or modify IHRL. Citing the U.N. Human Rights Committee, Kaye reminds us that we should not grant a margin of appreciation to states to define legitimate restrictions on speech, regardless of the robustness of their democratic institutions.[18] In the same vein, when Aswad analyzes a proposal to renegotiate and clarify some international treaties, she rejects the possibility: "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."[19] 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.[20]

When IHRL advocates discuss the divergences between U.N. and regional treaties, they disregard state consent as an essential factor. Interestingly, they acknowledge that all regional systems diverge to different extents from the U.N.[21] However, this does not matter. In his 2019 report, Kaye explains how these systems relate to the U.N. treaties. "Regional human rights norms cannot, in any event, be invoked to justify departure from international human rights protections."[22] Subpart II.B addresses how IHRL advocates try to iron out this contradiction

[14] *SR Report 2018*, *supra* note 2 ¶ 10.

[15] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8, at 34.

[16] Strossen, *supra* note 4, at 355-356; *Id.*, at 38.

[17] *SR Report 2018*, *supra* note 2, ¶ 41

[18] David Kaye (Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, No. A/74/486 (Oct. 2019) [hereinafter *SR Report 2019*] ¶ 27.

[19] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8, at 61.

[20] Aswad, *To Protect Freedom of Expression*, *supra* note 11, at 648-649.

[21] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8, at 44-45.

[22] *SR Report 2018*, *supra* note 2, ¶ 26

between state consent as the source of the legitimacy of IHRL and IHRL's indifference toward state consent.

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

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

Often scholars and advocates express concern that IHRL is too vague and internally contradictory to guide corporate policies effectively.[23] This concern finds two types of responses. Some scholars argue that IHRL provides clear enough guidance on many valuable aspects of content moderation. Evelyn Aswad and organizations like Article 19 have unpacked what IHRL demands from corporations. David Kaye's report on hate speech as Special Rapporteur offered well-established principles that companies would need to follow. However, like any other legal system, IHRL has areas of indeterminacy even in its most granular versions, and its most well-established principles admit exceptions.

Hate speech bans provide good examples. 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)."[24] The same report states that "Laws 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."[25] However, Kaye immediately adds that "the application of any such restriction under international human rights law should involve the evaluation of the six factors noted in the Rabat Plan of Action."[26] So even prohibitions that seem clear enough to provide guidance to companies on highly controversial questions admit exceptions under vague circumstances.

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.[27] However, Kaye clarifies that "Across a range of ills that may have a more pronounced

---

[23] Amal Clooney & Philippa Webb, *The Right to Insult in International Law*, 48 COLUM. HUM. RTS. L. REV. 1 (2017); evelyn douek, *The Limits of International Law in Content Moderation*, 6 UC IRVINE J. INT'L, TRANSNAT'L, & COMPAR. L. 37 (2021); Brenda Dvoskin, *International Human Rights Law Is Not Enough to Fix Content Moderation's Legitimacy Crisis*, BERKMAN KLEIN MEDIUM COLLECTION (Sep. 16, 2020), https://medium.com/berkman-klein-center/international-human-rights-law-is-not-enough-to-fix-content-moderations-legitimacy-crisis-a80e3ed9abbd.

[24] *SR Report 2018*, *supra* note 2, ¶ 10.

[25] *Id.*, ¶ 22.

[26] *Id.*

[27] *Side-stepping rights: Regulating speech by contract*, ARTICLE 19 1, 19 (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB.pdf ("[Facebook's] criteria are still far broader than those permitted under international law. For instance, 'attack' is broadly defined as encompassing 'violent speech,' 'dehumanising statements' or 'statements of inferiority' *without making any reference to either the intent of the speaker to incite others to take action, or the likelihood of a specific type of harm occurring as a result of the speech*

impact in digital space than they might offline—such as misogynist or homophobic harassment designed to silence women and sexual minorities, or incitement to violence of all sorts—human rights law would not deprive companies of tools. To the contrary, 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."[28] 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.

Another type of response is to regard that indeterminacy as a feature. Because IHRL is (of course, like any legal system) indeterminate, it would not impose answers. Instead, it would offer two possibilities: a margin of maneuver for companies to choose their ethos;[29] and a process for actors to be in conversation and reach those answers.[30] It is unclear how IHRL would facilitate conversations between different actors. If actors are missing in the discussion, it is more likely due to power differentials and a lack of appropriate institutions for that dialogue to occur, not because these actors need to find a common language to understand each other. In any event, there is a tension between arguing that IHRL offers standards that companies can voluntarily adhere to and asserting 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.

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

IHRL is proposed as common ground. Instead of highlighting an agreement between states, this justification emphasizes that IHRL is a minimum shared agreement between communities across the world. IHRL gives "a common language" or least common denominator. It provides a baseline for (unidentified) actors to be in conversation and reach more granular agreements.

---

at issue. The examples given suggest that many different types of legitimate speech are likely to be removed") (emphasis added).

[28] *SR Report 2018*, *supra* note 2, ¶ 43.

[29] *See e.g. SR Report 2019*, *supra* note 18, ¶ 48 ("When company rules differ from international standards, the companies should give a reasoned explanation of the policy difference in advance, in a way that articulates the variation. For example, were a company to decide to prohibit the use of a derogatory term to refer to a national, racial or religious group—which, on its own, would not be subject to restriction under human rights law—it should clarify its decision in accordance with human rights law."); SR *Report 2018*, *supra* note 2, ¶ 43 ("Yet human rights law is not so inflexible or dogmatic that it requires companies to permit expression that would undermine the rights of others or the ability of States to protect legitimate national security or public order interests. Across a range of ills that may have more pronounced impact in digital space than they might offline—such as misogynist or homophobic harassment designed to silence women and sexual minorities, or incitement to violence of all sorts—human rights law would not deprive companies of tools. To the contrary, 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.").

[30] 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, 967 (2020) ("a human rights-based approach provides platforms with a common conceptual language to identify the impact of their moderation rules, processes and procedures in different contexts and to explain, discuss, and justify their moderation decisions in an open and transparent manner"); *id.*, at 968 ("Importantly, international human rights law does not always dictate a specific or uniform outcome, but it provides a framework and vocabulary for platforms to assess whether the human rights impacts of their moderation systems are justifiable …").

At the same time, U.S. scholars argue that IHRL would bring these rules closer to the First Amendment doctrine and would be the most effective framework to fight increasingly expansive corporate speech codes.[31] For example, Nadine Strossen argues that it would be strategically advantageous for companies to adopt U.N. standards because they will be more readily accepted globally than the First Amendment doctrine, even though the two share many vital elements.[32]

Along similar lines, when U.N. authoritative interpretations conflict with decisions originating from European bodies, proponents of the IHRL project do not cast doubt that U.N. solutions should prevail. It is problematic to claim that a legal framework is open-ended and invites the search for collective answers while asserting that U.N. solutions should always preempt regional or local preferences.

II. The Ideal of Neutrality

Rather than highlighting these contradictions as necessarily problematic or trying to solve them, I want to focus on what these contradictions illuminate: that most pieces of the IHRL project, even if they contradict each other, share an ideal of neutral governance. Notably, they share a model of governance ruled by exogenous standards of what is good. These standards are objective and come from nowhere.[33] They do not reflect someone's point of view but are imagined as either agreed by every relevant actor or as self-evidently good.

At times, it sounds like neutrality would stem from process and everyone's participation, for instance, when IHRL proponents reference states' consent. Participation is a way of building neutrality that Sheila Jasanoff describes as building a view from everywhere. It resembles Habermas's model for making legitimate rules where legitimacy depends on everyone being heard.[34] However, when IHRL proponents invoke state consent, shared agreements, or a minimum global consensus, they imagine these processes have already taken place.

The most challenging question in online speech governance is who can govern it. The status quo answer—corporations—is deeply unsatisfying. But alternatives are not easy to build. The IHRL project posits that it is possible to create a governance system by answering only the question about what the fundamental norms are without answering the questions about who makes them and how. In the end, if rules are transparent, intelligible, and known, those who know them—experts—may apply them and clarify them when needed. Thus, the lack of answers on who should govern means in practice that experts may do so. This model is the opposite of how we may imagine a legislature, where a body of representatives sets rules based on what the public informs

[31] Strossen, *supra* note 4; Aswad, *To Protect Freedom of Expression*, *supra* note 11.

[32] Strossen, *supra* note 4, at 333 ("[n]otwithstanding 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.").

[33] Sheila Jasanoff, *The Practices of Objectivity in Regulatory Science*, *in* Social Knowledge in the Making (C. Camic, N. Gross, & M. Lamont eds. 2011) 307-337 (describing a mode of building objectivity in regulatory science that she describes as a view from nowhere).

[34] Jürgen Habermas, *Discourse Ethics: Notes on a Program of Philosophical Justification*, *in* Moral Consciousness and Communicative Action (C. Lenhardt and S. W. Nicholsen trans. 1990) [German, 1983].

them is good. Here, governance is conceived in the opposite direction: a small body informs the public what is good.

A. Self-evidently good

The organization Article 19 produced a policy brief in 2018 describing the steps social media companies should take to comply with international freedom of expression standards.[35] At no point do they explain why companies should respect these standards. When analyzing platforms' internal hate speech rules, they find that platforms' rules usually restrict speech that IHRL protects.[36] According to their account, the reasons that drive this lower speech protection are commercial: These rules enable platforms to grow their users' base and better accommodate advertisers' interests.[37] Here, there are two sides: rules driven by business interests and rules that favor the public interest. There is no question that IHRL and the public interest are equivalent.

Many agree that human rights are a good thing. It may be that the concept is so ill-defined that it leaves space for all claims. In one sense, that is the case: so many (all?) interests may find an anchor in a human right: speech, safety, dignity, non-discrimination, equality. All sides of any debate about restrictions on speech will be defending a human right. The open texture of human rights and their aura of "good" may be the reason why Twitter can say something like this: "This is a global commitment, and while grounded in the United States Bill of Rights and the European Convention on Human Rights, it is informed by a number of additional sources including the members of our Trust and Safety Council, relationships with advocates and activists around the globe, and by works such as United Nations Principles on Business and Human Rights."[38] The idea behind it must be that all good documents must overlap, despite the actual tensions between all these sources.

IHRL includes all these rights but also some instruments that define appropriate ways to balance these rights. And here, people do not agree. Take the proportionality test enshrined in Article 19 of the ICCPR. The test itself is so thin that it cannot be contested: restrictions on speech must be adequate to satisfy their own goals. These common-sense requirements are sometimes elevated to an enlightened formula for policy design and method for actors to follow and reach shared conclusions. It is easy to agree on the test. But IHRL offers more than the test: it also provides IHRL-certified good balances. Those are the ones reached by U.N. agencies: especially the Human Rights Committee and the various U.N. Special Rapporteurs. Overall, Article 19 gives a very thin guide that may be self-evidently good, but it also comes with specific ways of balancing rights that IHRL brings to the content moderation space like Trojan horses.

The comparison between conversations about platforms adopting First Amendment doctrines and IHRL as default also illuminates how IHRL is discussed as uncontestably good. Most scholars who like the First Amendment are resistant to asking companies to apply it as default rules for

[35] *Side-Stepping Rights*, *supra* note 27.

[36] *Id.*, at 16.

[37] *Id.*

[38] *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.

content moderation.[39] Scholars acknowledge that if platforms hosted all First Amendment–protected speech, platforms would be worse for almost all users. However, similar concerns have not been raised by IHRL supporters even though if social media allowed all the speech that IHRL protects, users and companies would have to tolerate spam and all types of undesirable speech. IHRL has a non-contestable character 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 one's normative preferences.

B. Consent

Global consent is a clear expression of the ideal of neutrality: if everyone has consented, no one loses. One could argue that in this case, objectivity is constructed as a view from everywhere: every state participates in IHRL making, and it is that process of participation that validates the legitimacy of international law. However, when IHRL advocates reference states' consent, they imagine that the participatory procedure has already taken place.[40] What is left are standards for good speech regulation that can be implemented in different contexts.

When IHRL advocates face clear divergences from that global consensus, this does not seem to challenge the imagined neutrality of the project. Variations—especially between the U.N. and regional systems of human rights—are, instead, incorporated into a coherent design with the help of additional exogenous principles.

Aswad explains that inconsistencies between international and regional mechanisms do not render the U.N. human rights regime incoherent: "It simply means that the U.N. system provides more protections for speech than the regional systems."[41] IHRL advocates imagine a Kelsen-style pyramid of international law similar to the pyramid that organizes the hierarchy of constitutional law and other laws or federal and state law. While constitutions may have in theory a stronger claim to democratic credentials, it is unclear why the Human Rights Committee or the U.N.-appointed Special Rapporteurs would represent more foundational social commitments than regional human rights conventions.

IHRL advocates may find a basis for this pyramid in the UNGPs. The UNGPs only refer to U.N. treaties. Therefore, Aswad argues that companies must look at those treaties: Twitter's statement about 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 international human rights law rather than domestic laws (like the U.S. Bill of Rights) or regional law (such as the European Human Rights Convention)."[42] The following section digs deeper into the UNGPs as an element of neutral governance.

---

[39] Brenda Dvoskin Dannecker, *Representation without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, 67 VILL. L. REV (forthcoming 2022).

[40] Aswad, *To Protect Freedom of Expression, supra* note 11 (arguing that this participatory procedure should not be reopened to negotiate more granular rules more readily 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).

[41] *Id.*, at 642.

[42] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8, at 44.

The pyramid of international law sometimes allows experts to apply regional treaties without jeopardizing the system's coherence. According to this pyramid, U.N. treaties and their authoritative interpretations function as a floor of rights and preempt regional discrepancies except when regional treaties "expand" human rights. What does the work here is an apparently neutral idea about what rights companies ought to expand. I am not aware of any case in which the expansion of one right would bear no costs on other rights. For example, in regulating prior restrains to speech, the American Convention on Human Rights is more protective of speech than any other system.[43] But this protection of speech comes at the expense of less protection for other rights such as privacy or safety. One can only see regional systems as expanding rights if one already knows what rights should be expanded.

C. U.N. *Guiding Principles*

The United Nations Guiding Principles on Business and Human Rights are the most intriguing example of how the IHRL project is justified. IHRL advocates remind us that the HRC unanimously decided that companies should respect IHRL. A project that is framed as giving power back to the public responds to a small bureaucracy.

Evelyn Aswad's framing of companies' interests provides a specific example of how the IHRL project values neutrality. Aswad asks the question about how IHRL may accommodate companies' own right to freedom of expression. This is a common concern among U.S. scholars discussing the possibility of imposing legal duties on how companies should moderate content.[44] Her answer: "requiring social media platforms to have speech codes based on international human rights law standards would not necessarily violate the speech rights of corporations under international human rights law as they do not hold such rights."[45] Consideration for companies' rights is no longer necessary as, under IHRL, they simply do not exist. Further, the UNGPs set the opposite expectation: that business enterprises will respect the human rights of impacted individuals. The shift from a First Amendment framework to an IHRL framework, through the UNGPs, makes the cost of constraining companies' interest disappear.

D. A problem of implementation

The most common criticism of the IHRL project lies in its indeterminacy: IHRL is too vague and does not offer precise enough guidelines on how to address the novel questions posed by technological developments. This line of criticism can also be premised on an ideal of neutral governance. It does not need to challenge IHRL as standing for what is good. Instead, it frames the problem as one of implementation: how to apply what we already know is good.[46]

---

[43] Article 13 provides that "[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."

[44] *See e.g.* Daphne Keller, Who Do You Sue? State and Platform Hybrid Power Over Online Speech (Hoover Inst. Aegis Paper Series, Paper No. 1902, 2019).

[45] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8, at 41.

[46] Sander, *supra* note 30, at 969 ("Arguably the biggest challenge, however, resides in the *translation* of general

The challenge of implementation takes two different forms: IHRL must be implemented in different geographical contexts, and it must be translated to adapt to the affordances of social media platforms. On the first challenge, diversity and specifically geographical expertise must complement human rights expertise. For IHRL advocates, local differences mean that context needs to be considered in order to apply those global norms adequately. The challenge is to understand the local context. These differences are presented as a matter of expertise. One can hire an expert in the local context to ensure that the global standard is applied fairly.

Companies address this issue by creating partnerships with local experts through "trusted partner programs." These programs afford local civil society organizations a privileged channel to report content that should be taken down or to appeal decisions when content has been taken down incorrectly. Often companies will resort to their trusted partners to *understand* the local context of a protest or how a word is used in the country in order to apply their corporate rules accordingly. Local NGOs assist in monitoring social media and applying companies' community guidelines. These programs originate in companies' recognition that they operate across multiple and different contexts. But the programs make these differences a matter of implementation rather than disputes over how companies should regulate speech.

The Facebook Oversight Board (FOB) is also very 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. The FOB usually makes this assessment under its analysis of the "necessity" test under Article 19 of the ICCPR. Whether a rule is necessary to achieve a specific goal could be understood as a matter of normative preferences. The requirement is that the restriction in speech is proportional to the pursued objective. This implies answering questions such as how many false positives are tolerable to prevent certain harm or what level of efficiency deems the restriction "proportional." 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 some errors will be unavoidable. Whether the most expansive ban is proportional to the goal may depend on how a society values the protection of nudity online or how essential it considers the avoidance of nonconsensual distribution.

That is not the kind of local context that the FOB considers necessary to take into account. Rather, it is the local context that experts can use to implement Facebook's rule correctly.[47] In Case decision 2021-010-FB-UA concerning Facebook's decision to remove a post showing a video of a protester in Colombia criticizing the country's president and using a slur that discriminates against gay men, the FOB considered that Facebook had not taken the local context sufficiently into account. On the one hand, Facebook did not understand how the slur is used in Colombia with different meanings that are not always discriminatory. More importantly, eliminating the post was not proportional in the political context in Colombia, where protests

human rights principles into particular rules, processes and procedures tailored to the platform moderation context."); 970 ("The remainder of the Article takes up this final challenge, exploring some of the choices and hurdles that online platforms are likely to confront in attempting to *implement* their corporate responsibility to respect in the content moderation context."); 971 ("…*translation* from the State to the corporate context of platform moderation is likely to pose a number of challenges in practice") (emphasis added).

[47] *See* Case 2021-008-FB-FBR; Case 2021-010-FB-UA.

against the administration were taking place and social media played a significant role in sharing information about the protest. The FOB cited the U.N. Special Rapporteur on freedom of expression to justify allowing the discriminatory term to stay on the platform. The Special Rapporteur stated that "evaluation of context may lead to a decision to make an exception in some instances when the content must be protected as, for example, political speech."[48] At the end of the decision, the FOB adds, as in all other cases, that "[a]n independent research institute headquartered at the University of Gothenburg and drawing on a ream of over 50 social scientists on six continents, as well as more than 3,200 country experts from around the world, provided expertise on socio-political and cultural context." Overall, local context is something that experts need to see, understand, and explain to Facebook to ensure the correct application of its rules.

In addition, IHRL requires translation from the state to the corporate context. In some cases, the FOB has allowed Facebook to adopt rules that, in the FOB's view, directly contradict what IHRL prescribes. Even in these cases, the FOB has justified these rules as an appropriate implementation of human rights. A series of decisions concerned Facebook's ban on a list of racial slurs. In Case Decision 2021-002-FB-UA regarding Facebook's prohibition on content depicting blackface, the FOB upheld this ban even though if a state were to restrict speech in this same way, it would be violating its international obligations. However, the technical specificities of Facebook made these rules acceptable translations of IHRL. In Section III, I analyze these decisions in detail.

III. Neutrality and Participation

IHRL advocates argue that because IHRL is indeterminate, the project requires a collective conversation to find answers in those vague areas. Evelyn Aswad proposes that multi-stakeholder initiatives debate and decide what aims corporate restrictions on speech may legitimately pursue.[49] David Kaye insists on the importance of the input of local civil society organizations.[50] This section looks at how the IHRL project envisions that this conversation may occur and what IHRL itself may contribute to the dialogue.

On the one hand, IHRL may contribute to enabling this conversation because it requires that corporations and institutions create participatory mechanisms. Principles 17 and 18 of the UNGPs prescribe that corporations should carry out human rights due diligence, which should "[i]Involve meaningful consultation with potentially affected groups and other relevant stakeholders, as appropriate to the size of the business enterprise and the nature and context of the operation."

These proposals for multi-stakeholder conversations or consultation mechanisms envision another kind of neutral governance. Instead of relying on exogenous standards (what Jasanoff calls a viewpoint from nowhere), these initiatives aim at building a viewpoint from everywhere.[51] They require a process whose outcome is legitimate if all views have adequate representation.

On the other hand, IHRL may contribute substantively to these dialogues. IHRL demands transparency and public justifications for corporate practices. Transparency requirements would

[48] A/74/486, para. 47 (d) cited in Case 2021-010-FB-UA, page 13.

[49] Aswad, *The Future of Freedom of Expression Online*, *supra* note 8.

[50] Blayne Haggart & Clara Iglesias Keller, *Democratic legitimacy in global platform governance*, 45 Telecomm. Pol'y 1, 11 (2021) (analyzing the role of civil society organizations in Kaye's proposal).

[51] Jasanoff, *supra* note 33.

enable more informed and robust conversations. However, 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 language" to be in conversation. The test set out by Article 19 of the ICCPR could serve as a reasoning process to guide the conversation. The test itself, however, stripped of its requirements that a government body acts and that restrictions on speech advance specific goals, is so thin that it is unclear how it would play a role in channeling debates.

It could be that the common language includes more than this adapted version of Article 19. It could be that once IHRL is accepted as the appropriate framework, all documents produced by U.N. agencies and experts and perhaps from other sources will gain new authority and persuasiveness. New actors who know or learn how to phrase their demands in the terms of these documents may be able to join the conversation, relying or not on institutional mechanisms for participation. These actors may find new ways to disrupt neutrality and manage to intervene in spaces that were previously out of their reach.

A. Top-down participation: between Dworkin and Habermas

By top-down participation, I refer to institutional opportunities created by those in power to enable or encourage the participation of new actors. Instead of relying on self-validating standards, neutrality is achieved by procedural safeguards that ensure the representation of all viewpoints. The view from nowhere discussed in Section II is applied by the Herculean experts that Dworkin envisioned who can find and correctly translate principles into decisions. This view from everywhere is achieved by Habermasian norm-setting procedures.

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.[52] Are these mechanisms gaining new force thanks to the IHRL project?

The Facebook Oversight Board has shown in different opportunities its interest in mechanisms for civil society organizations to participate. It provides opportunities for the public to submit comments in all cases, in a similar fashion to amici curiae.[53] In some instances, it has established rudimentary forms of dialogue with civil society organizations, directly citing reports or addressing concerns expressed by civil society.[54] And it has valued Facebook's efforts to engage civil society. In some cases, it has framed these mechanisms as IHRL duties and has recommended that Facebook extend the reach of its engagement efforts.[55]

In Case Decision 2020-006-FB-FBR concerning a post containing inaccurate information about COVID-19 treatments, the FOB recommended that "Facebook should conduct a human

[52] Matthias Kettemann & Wolfgang Schulz, *Setting Rules for 2.7 Billion. A (First) Look into Facebook's Norm-Making System: Results of a Pilot Study*, Working Papers of the Hans-Bredow-Institut (2020); Dvoskin, *Representation without Elections*, *supra* note 39.

[53] *See e.g.* Case Decision 2020-006-FB-FBR, 15 (stating that the recommendations to Facebook drew on public comments the Board received).

[54] *See e.g.* Case 2021-FB-UA, 11 (citing to a report elaborated by the NGO Media Monitoring Africa).

[55] *See e.g.* Case 2021-FB-UA, 10 ("To meet its human rights responsibilities when developing and reviewing policies, including the slur list, Facebook should consult potentially affected groups and other relevant stakeholders, including human rights experts.").

rights impact assessment with relevant stakeholders as part of its process of rule modification" per Principles 18 and 19 of the UNGPs.[56] In Case Decision 2021-006-IG-UA concerning a post discussing the solitary confinement of Abdullah Öcalan, the FOB recommended that Facebook "ensure meaningful stakeholder engagement" to review its policy on dangerous individuals and organizations, including engagement through a public call for inputs.[57] Notice that in this case, the FOB did not refer to any specific IHRL obligation. Perhaps the FOB can make this recommendation in a similarly convincing fashion regardless of the reference to semi-legal standards. It is intriguing why it made this recommendation in these cases and not in others.

This interest in stakeholder engagement may evolve into more precise standards for stakeholder consultation. For example, in Case Decision 2021-011-FB-UA regarding the use of a racial slur in South Africa, not only did the FOB appreciate Facebook's consultation with external stakeholders to draft an exception to its hate speech policy for insults when used self-referentially, but it also considered relevant that the external stakeholders represented diverse geographical regions.[58]

In other cases, the FOB hinted timidly that consultation with civil society organizations could lead the FOB to defer to Facebook's policies and even 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 determination because Facebook had consulted with regional and linguistic experts and had worked with a local partner to identify and adjudicate the content. The Board was interested in engagement with these actors, even though IHRL did not seem to play a significant role in shaping or justifying this interest. In Case Decision 2021-002-FB-UA regarding Facebook's ban on content depicting blackface, the FOB 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."[59] The majority considered that this stakeholder consultation was in line with "international standards for human rights due diligence" and cited to Principles 17(c) and 18(b) of the UNGPs.

This decision shows that despite the interest in stakeholder engagement initiatives, the view from nowhere has so far prevailed, at least in how the FOB is deploying the IHRL project. The Oversight Board decided to uphold a general ban on blackface, even though, according to the Board, the same prohibition would be incompatible with IHRL if adopted by a state because this speech does not necessarily constitute an incitement to violence. The Board appreciated the process to engage stakeholders, although a minority considered that Facebook had provided insufficient information on the extent of stakeholder engagement in countries that celebrate the Black Pete tradition, at issue in this case.[60] However, when justifying the decision to uphold the ban on content depicting blackface, the FOB did not reference this process.

[56] Case 2020-005-FB-FBR, 13.

[57] Case 2021-006-IG-UA.

[58] Case 2021-011-FB-UA, 7.

[59] Case 2021-002-FB-UA, 11.

[60] *Id*, at 11.

Instead, the FOB decided that the divergence was justified because many experts had found that images depicting blackface are discriminatory and harmful. It did not consider the ban to be a departure from IHRL: "The majority found Facebook followed international guidance and met its human rights responsibilities in this case."[61] To justify this finding, the Board relied on the fact that "[n]umerous human rights mechanisms have found the portrayal of Zwarte Piet to be a harmful stereotype."[62] But not any harmful stereotype would satisfy the Board. The FOB cited many reports and concluded that these expert findings provided "sufficient evidence of *objective* harm to individuals' rights to distinguish this rule from one that seeks to insulate people from *subjective* offense."[63] 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 that determine an objective reason to make this departure from IHRL compatible with IHRL.

Another example in the same direction comes from Case decision 2021-011-FB-UA concerning the use of a racial slur in South Africa. 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.

This case offers an additional glimpse into how the Board divides tasks among U.N. experts and civil society organizations. On the one hand, the FOB claims that banning racial slurs is incompatible with IHRL. Still, Facebook can do it because the U.N. "Special Rapporteur indicates that entities engaged in content moderation like Facebook can regulate such speech."[64] On the other hand, the FOB stated that Facebook should consult affected groups and human rights experts, as it did in the case, "to meet its human rights responsibilities when developing and reviewing policies."[65] Notice that this consultation aims to understand the meaning of the terms in the context in which they are used. Participation is thought of as an implementation mechanism, while the rule-making process is kept in the hands of a small group of experts.

Other actors walking the line between exogenous standards and participatory procedures are international human rights organizations. Agencies such as UNESCO, the Special Rapporteurs, and the Inter-American Commission on Human Rights are beginning to regard setting up norms for online speech regulation as part of their role.[66] Of particular relevance here is the initiative that the Inter-American Commission on Human Rights has recently launched to hold a multi-stakeholder dialogue on how to make content moderation policies "compatible" with IHRL

[61] *Id,* at 14.

[62] *Id*.

[63] Case 2021-002-FB-UA, 16 (emphasis added).

[64] Case decision 2021-011-FB-UA, 14

[65] *Id.,* 10.

[66] *See e.g.* Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda", Office of the Special Rapporteur for Freedom of Expression, 2017.

standards.[67] The initiative contemplates multiple opportunities for seeking input from experts, the general public, other regional and international endeavors, and multilateral initiatives such as the Open Government Partnership. As the initiative develops, it could be a significant effort to set authoritative standards for content moderation built outside companies, borrowing legitimacy from participatory procedures and using the IHRL framework to strengthen the authority and persuasiveness of the substantive outcomes.

B. Bottom-up participation: IHRL disrupting neutrality

Institutionalized opportunities are important but never enough. A healthy system of governance requires both institutionalized and spontaneous actors.[68] Spontaneous actors can change such institutions, participate from the outside, and challenge institutional contours and operations. Traditionally, the electorate is the central actor that plays that role. Can IHRL empower spontaneous actors in the world of online speech governance? I believe the answer is yes.

On the one hand, many (all?) actors involved in designing content moderation rules and practice claim to be doing human rights work. Some organizations are more specific about what their support for IHRL requires. Jillian York and Corynne McSherry demand that companies align their policies with human rights norms and recommend that "censorship must be rare."[69] Access Now and Article 19 have produced a detailed report unpacking what rules platforms need to adopt in order to respect IHRL (spoiler: they need to protect more speech).[70] Others expressly acknowledge that they have turned to the IHRL project because they have not been able to join the conversation successfully so far.[71] And organizations like the Center for Democracy and Technology, traditionally detached from discussions on how private actors moderate content, have recently joined the conversation advocating for an adapted version of IHRL.[72]

Organizations in Latin America have recently produced a policy brief articulating what IHRL would require, in their view, from companies.[73] Speaking the IHRL language may turn up the volume of the demands of organizations in this region.

---

[67] Inter-American Commission on Human Rights, *Americas Dialogue on Freedom of Expression Online*, AMERICAS DIALOGUE, https://www.americasdialogue.org/.

[68] Gunther Teubner, *Societal Constitutionalism: Alternatives to State-Centered Constitutional Theory*, Storrs Lectures 20003/04 Yale Law School, 1, 23.

[69] Jillian C. York & Corynne McSherry, Content Moderation is Broken. Let Us Count the Ways, ELECTRONIC FRONTIER FOUNDATION DEEPLINKS BLOG (Apr. 29, 2019), https://www.eff.org/deeplinks/2019/04/content-moderation-broken-let-us-countways.

[70] *Protecting Free Expression in the Era of Online Content Moderation*, ACCESS NOW (May 2019), https://www.accessnow.org/cms/assets/uploads/2019/05/AccessNow-Preliminary-Recommendations-On-Content-Moderation-and-Facebooks-Planned-Oversight-Board.pdf; *Side-Stepping Rights, supra* note 27.

[71] Strossen, *supra* note 4.

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

[73] Observacom et al., *Estándares para una regulación democrática de las grandes plataformas que garantice la libertad de expresión en línea y una Internet libre y abierta*, OBSERVACOM (July 2020), https://www.observacom.org/wp-content/uploads/2020/09/Estandares-para-una-regulacion-democratica-de-las-

As it emerges from this array of actors, IHRL more easily captures the preferences of advocates for robust protections of speech and weaker protections for other rights. It may be that these organizations can make their positions more persuasive and influential. However, the newly empowered positions from some advocates may be a small benefit compared to the risks that this paper has highlighted. Because articulating demands in this language makes them sound objective and neutral and because the FOB is in an exceptional position of power to articulate these preferences as neutral,[74] the risk is that we will end up with uncontestable speech principles and a small bureaucracy tasked with enforcing them.

CONCLUSION

International human rights law promises to rein in corporate power on behalf of the public interest. Instead of focusing on restructuring institutional power, this project proposes to achieve its goals by offering substantial standards that are global, shared, and good. Where it encounters unavoidable ambiguities and disagreements, the project either borrows additional exogenous principles to resolve them or turns to expert knowledge that can provide the correct answer. The risk is to deny that conflict exists and to believe that answers that proceed from a small bureaucracy adequately represent the public interest.

grandes-plataformas.pdf.

[74] However, the FOB's history is very brief and we could have the impression that IHRL, as applied by the FOB, is very speech protective only because until now they have mostly reviewed decisions involving speech that had been taken down.