Applying International Human Rights Law for Use by Facebook

Michael Lwin[†]

In recent years, social media platforms have been beset with hate speech, misinformation, disinformation, incitement of violence, and other content that cause real-world harm. Social media companies, focusing solely on profit-maximization and user-engagement, have been largely asleep at the wheel during outbreaks of violence in countries such as Myanmar, Sri Lanka, New Zealand, and India–events all linked in some way to online content. When social media companies began trying to reduce harmful content, they made tweaks: incremental, non-transparent, and often inconsistent changes to their moderation rules. To build a more effective and consistent system, some international lawyers have suggested that social media companies adopt international human rights law (IHRL)–especially the International Covenant for Civil and Political Rights (ICCPR)–as a unified source for content moderation rules. However, IHRL was written and ratified for use by states, not private companies. Moreover, IHRL emerged long before the Internet and social media were widespread. IHRL must therefore be interpreted and adapted for this new purpose. As a first step towards honing and refining its application, this article proposes a framework for the use of IHRL by social media companies.

[†] Managing Director, Koe Koe Tech. I would like to give my sincere thanks to Susan Benesch, Matthew Bugher, Zoe Darmé, Khoe Reh, Andrew Smith, Brent Harris, Abigail Bridgman, Thomas Kadri, Thomas Hughes, Katherine Bond, Noah Feldman, Sarah Oh, Matthew Smith, Wai Yan, Benjamin Staubli, Winsandar Soe, and Thant Sin for their comments and insights informing this essay. I would also like to thank Rachel Brown, Talya Lockman-Fine, Myat Su San, Rohan Subramanian, Frances O'Morchoe, Andrew Santana, and John Willis for their hard work editing this essay.

I. Introduction

> We are making rules up.
>
> Facebook Policy Team Member[1]

The slow response by social media companies to halt the spread of hate speech, misinformation, disinformation, incitement of violence, and other attempts to invoke harm on their platforms is well known by now. These include the incitement of mob violence on WhatsApp and Facebook in India,[2] the Christchurch shooter livestream on YouTube, Reddit, and Facebook,[3] and the relentless building of support for mass atrocities in Myanmar.[4]

The prevailing paradigm of techno-libertarianism, which focuses solely on profit-maximization and active user retention and engagement, leads to real harm. During the early years of social media, the ties between platforms and violence were easier to ignore, and the people who ran most of these companies did so. Take Facebook as an example. There were no international lawyers or cultural anthropologists at the top levels of Facebook, which helped to explain the company's initial lack of response and subsequent fumbling reactions to freedom of expression and incitement to violence issues on the platform. In response to sustained criticism,[5] Facebook has since hired people from the international development space to create their Oversight Board. These included former UN Peacekeeping Officer, Zoe Darmé,[6] the ICC's Abigail Bridgman,[7]

---

[1] MATTHIAS C. KETTEMANN & WOLFGANG SCHULZ, LEIBNIZ INSTITUTE FOR MEDIA RESEARCH, SETTING RULES FOR 2.7 BILLION: A (FIRST) LOOK INTO FACEBOOK'S NORM-MAKING SYSTEM: RESULTS OF A PILOT STUDY, 28 (Jan. 2020) https://leibniz-hbi.de/uploads/media/Publikationen/cms/media/5pz9hwo_AP_WiP001InsideFacebook.pdf.

[2] Amanda Taub & Max Fisher, *Where Countries Are Tinderboxes and Facebook Is a Match*, N.Y. TIMES (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html [https://perma.cc/N4RL-WFJ5].

[3] Kevin Roose, *A Mass Murder of, and for, the Internet*, N.Y. TIMES (Mar. 15, 2019), https://www.nytimes.com/2019/03/15/technology/facebook-youtube-christchurch-shooting.html [https://perma.cc/9EFQ-QNVT].

[4] Steve Stecklow, *Why Facebook Is Losing the War on Hate Speech in Myanmar*, REUTERS (Aug. 15, 2018) https://www.reuters.com/investigates/special-report/myanmar-facebook-hate [https://perma.cc/9HMN-328M].

[5] *See, e.g.*, Kevin Roose & Paul Mozur, *Zuckerberg Was Called Out Over Myanmar Violence. Here's His Apology*, N.Y. TIMES (Apr. 9, 2018), https://www.nytimes.com/2018/04/09/business/facebook-myanmar-zuckerberg.html [https://perma.cc/Y67T-WAWQ]; George Soros, *Remove Zuckerberg and Sandberg from Their Posts*, FINANCIAL TIMES (Feb. 17, 2020), https://www.ft.com/content/88f6875a-519d-11ea-90ad-25e377c0ee1f [https://perma.cc/M7LD-LDZL].

[6] Jen Patja Howell, *The Lawfare Podcast: Zoe Darmé on the Facebook Oversight Board*, LAWFARE (Nov. 14, 2019), https://www.lawfareblog.com/lawfare-podcast-zoe-darme-facebook-oversight-board [https://perma.cc/JWK9-2ZCK]. Note, however, that Zoe Darmé and Sarah Oh–the latter of whom was the leader of strategic response at Facebook–are no longer with Facebook. The author has recently been informed that Facebook is reducing staffing for policy teams.

[7] Abigail Bridgman is no longer with Facebook; she has joined the Oversight Board as a Case Selection Manager as of this writing.

and the International Center for Transnational Justice's Miranda Sissons.[8] Most recently, Facebook hired Thomas Hughes, former Executive Director of INGO ARTICLE 19, to be Chief Administrator of its pending Oversight Board.[9] It is unclear, however, whether all of these new hires and expenses will actually result in tangible and positive change at Facebook or just amount to another elaborate corporate greenwashing scheme for business as usual.[10]

Matthias C. Kettemann and Wolfgang Schulz of the Leibniz Institute for Media Research were permitted by Facebook to conduct a "pilot study into the private order of communication" at Facebook and "its policy development process."[11] Their study shows how Facebook's platform is not grounded in any one national legal order. Facebook's policies are influenced by competing interests, but overall by the preferences of leaders such as Mark Zuckerberg, Sheryl Sandberg, Monika Bickert (VP of Global Policy Management),[12] and Joel Kaplan (VP for US Public Policy).[13] Kettemann and Schulz, in their observations of Facebook employees, noted that even those employees with backgrounds in human rights failed to refer to concrete human rights norms during working group discussions or stakeholder engagements.[14] Facebook employees themselves acknowledge that "we are making rules up."[15] There exists no formal framework or procedures for content moderation decisions. As Kettemann and Schulz note, "[i]n any normative-social setting it holds true that, if the outcome of a procedure might be–for any reason–not intrinsically legitimate, then the proceduralization can increase its legitimacy and make a normative change amenable to those not agreeing with the

---

[8] Joshua Brustein, *Facebook's First Human Rights Chief Confronts Its Past Sins*, BLOOMERG (Jan. 28, 2020) https://www.bloomberg.com/news/articles/2020-01-28/facebook-s-first-human-rights-chief-seeks-to-tame-digital-hate [https://perma.cc/GQ5K-DLC5].

[9] Brent Harris, *Preparing the Way Forward for Facebook's Oversight Board*, FACEBOOK (Jan. 28, 2020), https://about.fb.com/news/2020/01/facebooks-oversight-board.

[10] *See,* Julia Carrie Wong, *Will Facebook's New Oversight Board Be a Radical Shift or a Reputational Shield?*, THE GUARDIAN (May 2, 2020), https://www.theguardian.com/technology/2020/may/07/will-facebooks-new-oversight-board-be-a-radical-shift-or-a-reputational-shield [https://perma.cc/D3QG-K7TS] (quoting Siva Vaidhyanathan) ("'I wish I could say that the Facebook review board was cosmetic, but I'm not even sure that it's that deep,' said Siva Vaidhyanathan, a professor of media studies at the University of Virginia and author of a book on Facebook. 'If Facebook really wanted to take outside criticism seriously at any point in the past decade, it could have taken human rights activists seriously about problems in Myanmar; it could have taken journalists seriously about problems in the Philippines; it could have taken legal scholars seriously about the way it deals with harassment; and it could have taken social media scholars seriously about the ways that it undermines democracy in India and Brazil. But it didn't. This is greenwashing.'").

[11] KETTEMANN & SCHULZ, *supra* note 1, at 5.

[12] KETTEMANN & SCHULZ, *supra* note 1, at 28-29.

[13] *See, e.g.*, Ben Smith, *What's Facebook's Deal With Donald Trump?*, N.Y. TIMES (Jun. 21, 2020), https://www.nytimes.com/2020/06/21/business/media/facebook-donald-trump-mark-zuckerberg.html [https://perma.cc/B6RT-QPQV].

[14] KETTEMANN & SCHULZ, *supra* note 1, at 32.

[15] KETTEMANN & SCHULZ, *supra* note 1, at 28.

particular policy outcome as well."[16] I strongly agree with Kettemann and Schulz on this point. In this essay, I argue that the framework I propose can act as the "proceduralization" necessary to increase both Facebook's and the Oversight Board's legitimacy.

In seeking a single consistent and effective source of content moderation rules, scholars have suggested that social media companies turn to international human rights law (IHRL), especially certain instruments on freedom of speech and speech regulation.[17] However, IHRL was written for use by states, not private enterprises. Moreover, its language can often be contradictory. For example, Article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) requires states to "declare an offence punishable by law all dissemination of ideas based on racial superiority or hatred, incitement to racial discrimination, as well as all acts of violence or incitement to such acts against any race or group of persons of another colour or ethnic origin."[18] Article 19 of the International Covenant on Civil and Political Rights (ICCPR), on the other hand, only outlines *optional* state prohibitions, which must comply with a three-part test of legality, legitimacy, and necessity, thereby conflicting with ICERD's Article 4 *requirements* for state prohibitions. Both of these articles further conflict with the narrow standard for required state prohibitions under ICCPR Article 20.[19]

IHRL institutions and observers are themselves aware of these inconsistencies. The ICERD Committee's General Recommendation No. 35–adopted in 2013–claws back the expansive language of ICERD Article 4, recommending "that the criminalization of forms of racist expression should be reserved for serious cases, to be proven beyond reasonable doubt, while less serious cases should be addressed by means other than criminal law, taking into account, inter alia, the nature and extent of the impact on targeted persons and groups. The application of criminal sanctions should be governed by principles of legality, proportionality and necessity,"[20] with the last clause being a clear reference to ICCPR Article 19(3). The ICERD committee gets its recommendation powers from the ICERD treaty.[21] The Special Rapporteur for Freedom of Expression,

[16] KETTEMANN & SCHULZ, *supra* note 1, at 28.

[17] *See, e.g.*, Evelyn Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26-70 (2018); ARTICLE 19, *Side-stepping Rights: Regulating Speech by Contract* (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf; David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018), https://digitallibrary.un.org/record/1631686/usage?ln=en.

[18] International Convention on the Elimination of All Forms of Racial Discrimination (ICERD) art. 4, Dec. 21, 1965, T.I.A.S. 94-1120, 660 U.N.T.S. 195.

[19] International Covenant on Civil and Political Rights (ICCPR) art. 19 and 20, Dec. 16, 1966, 999, T.I.A.S. 92-908, U.N.T.S. 171.

[20] UN Committee on the Elimination of Racial Discrimination, *General recommendation No. 35: Combatting Racist Hate Speech*, para. 12, CERD/C/GC/35 (Sept. 26, 2013), https://www.refworld.org/docid/53f457db4.html.

[21] ICERD, *supra* note 18, at *art.* 9(2).

David Kaye, claims that the ICERD committee "explained that the conditions defined in article 19 of the International Covenant on Civil and Political Rights also apply to restrictions under article 4 of the International Convention on the Elimination of All Forms of Racial Discrimination."[22] IHRL institutions often rely on their special rapporteurs and committees to smooth over or issue clarifying interpretations of the language in multilateral treaties.

Further, decisions by human rights tribunals are also often flawed and inconsistent. It is conceivable that the Oversight Board could adopt IHRL, in whole or in part (more likely in part and with adaptations), via one or more binding content decisions into the Oversight Board's jurisprudence. If done too broadly or unthoughtfully, such action may prompt (i) human rights lawyers to demand formal adoption of regional and domestic tribunal decisions, and (ii) countries to compel the Oversight Board to adopt domestic legislation, much of which tends to be overbroad, excessively restrictive, and overly favourable towards state interpretations of expression and incitement.

IHRL itself can be an unclear and inconsistent set of laws and rules. How can Facebook and its Oversight Board be expected to come up with clear and consistent content moderation rules while at the same time respecting IHRL? Is it possible for Facebook and the Oversight Board to avoid acting as judge, jury, and executioner? The answer is yes. This essay attempts to apply IHRL to social media companies in a way that is actionable not only for the companies themselves, but also the entities that regulate their content moderation decisions such as the Oversight Board that Facebook has promised to establish[23] as well as the Social Media Councils that Article 19 has proposed.[24] Although the framework is intended for use by any platform, this essay will focus on Facebook and its Oversight Board given the fact that the Board has been charged with adopting new jurisprudence.

II. Why Should IHRL Apply to Social Media Companies?

A. Social Media Companies Lack Legitimacy

The new Oversight Board has an opportunity to address the deficit of transparency and legitimacy surrounding Facebook's current content moderation rules and processes. However, the Board, as currently set up, risks perpetuating these problems. The Oversight Board's legitimacy

[22] David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, para. 16, U.N. Doc. A/74/486 (Oct. 9, 2019), https://www.ohchr.org/Documents/Issues/Opinion/A_74_486.pdf.

[23] Harris, *supra* note 9.

[24] ARTICLE 19, *Social Media Councils: Consultation* (Jun. 11, 2019), https://www.article19.org/resources/social-media-councils-consultation.

comes from Facebook itself. The Oversight Board has its own charter[25] and bylaws,[26] which were drafted by Facebook. It is composed of 40 members, including 4 co-chairs, who are selected by Facebook.[27] Unsurprisingly, many observers have noted that the creation of the Oversight Board, however well-intentioned, may have been a cover for Facebook to continue business as usual.[28] This conflict-of-interest issue persists in the tests that Facebook says it has developed. In 2019, Monika Bickert, Vice President of Global Policy Management and Chair of the Product Policy Forum at Facebook, said that their main goal was less about introducing new rules but rather providing "clarity".[29] Along this vein, Facebook has recently changed certain values in the preamble of its Community Standards. These values now comprise voice, authenticity, safety, privacy, and dignity.

Voice is the "paramount" value. The "goal of [Facebook's] Community Standards is to create a place for expression and give people voice. Building community and bringing the world closer together depends on people's ability to share diverse views, experiences, ideas and information."[30] It appears that "voice" is largely equivalent to the concept of

---

[25] Brent Harris, *Establishing Structure and Governance for an Independent Oversight Board*, FACEBOOK (Sept. 17, 2019), https://about.fb.com/news/2019/09/oversight-board-structure.

[26] Harris, *supra* note 9.

[27] *Oversight Board Bylaws* § 1.1.2, FACEBOOK (Jan. 2020) https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf ("Facebook will select the initial co-chairs.").

[28] *See, e.g.*, Evelyn Douek, *How Much Power Did Facebook Give Its Oversight Board?*, LAWFARE (Sep. 25, 2019), https://www.lawfareblog.com/how-much-power-did-facebook-give-its-oversight-board [https://perma.cc/V4JN-TN54] ("I have long been arguing that if the board's "subject matter jurisdiction" (that is, the range of Facebook decisions that it is empowered to review) is too narrow, its legitimacy will be undermined. For example, if the board can review only cases where Facebook decides to take a post down and has no power to review algorithmic ranking decisions, Facebook can avoid having the board pronounce on cases it would prefer not to be overruled on by simply downranking a troubling post so that no one sees it without formally triggering the board's jurisdiction. Ad policies, and especially political ad decisions, are critical decisions about political discourse—yet they can be opaque and inconsistent. There is no reason these decisions should not also be the subject of independent review."); Shirin Ghaffary, *Here's How Facebook Plans to Make Final Decisions About Controversial Content It's Taken Down*, VOX (Jan. 28, 2020) https://www.vox.com/2020/1/28/21112253/facebook-content-moderation-system-supreme-court-oversight-board [https://perma.cc/U6VB-TKCB] ("Several experts Recode spoke with called the updates a step in the right direction for greater transparency, but say that the project's success will depend on how much the company actually listens to this new governing body. Under the proposed rules, Facebook will be forced to follow the board's decisions when it rules that the company should not have taken down content. But for broader policy decisions, Facebook will only take guidance — not mandates — from the board."); Casey Newton, *Facebook is Putting a Surprising Restriction on Its Independent Oversight Board*, THE VERGE (Jan. 30, 2020), https://www.theverge.com/interface/2020/1/30/21113273/facebook-oversight-board-jurisdiction-bylaws-restrictions [https://perma.cc/NUL7-4VM3]. ("I'm less certain the board will have a say here. It will have the authority to remove (or leave standing) individual pieces of content, as well as issue policy advisory opinions. Key word: advisory. And while an opinion by the board that Facebook should fact-check political ads would have some weight — and could provide political cover for Facebook to reverse course, should it decide it wants to — ultimately the decision will likely still remain with Zuckerberg.").

[29] *See* KETTEMANN & SCHULZ, *supra* note 1, at 17 (quoting Monika Bickert).

[30] *Community Standards,* FACEBOOK*, https://www.facebook.com/communitystandards* (last visited July 21, 2020).

freedom of expression. When Facebook limits expression or voice, it does so by appealing to the other stated values of authenticity, safety, privacy, and dignity.[31] As Kettemann and Schulz note, "[i]n light of these values, Facebook professes to not wanting 'people using Facebook to misrepresent who they are or what they're doing,' trying to minimize expression 'that threatens people [and] has the potential to intimidate, exclude or silence others,' with the aim of protecting 'personal privacy and information' and ensuring that users 'respect the dignity of others and not harass or degrade others.'"[32] Voice can also be enhanced by notions of newsworthiness and public interest.[33] Thus, even in cases where voice may be restricted by the other values, it can still win out if it is augmented by newsworthiness and public interest. Facebook handles this through a balancing test that involves "weighing the public interest value against the risk of harm, and [looking] to international human rights standards to make these judgments."[34]

The central issue, however, is that "risk of harm," "newsworthiness," "public interest," and "international human rights standards" remain undefined in Facebook's Community Standards, public statements, postings, as well as Kettemann and Schulz's observations.[35] While Facebook does publish minutes of its Product Policy Forum meetings, meeting participants do not apply structured frameworks or articulate the actual quasi-legal reasoning used to reach their conclusions.[36] Furthermore, the balancing test is hard to square with how Facebook conducts policy change processes. When conducting such processes, Facebook asks "[d]o folks have concerns?" Here, the "folks" refer to "civil society organizations, activist groups, and thought leaders, in such areas as digital and civil rights, anti-discrimination, free speech, and human rights."[37] From the civil society and human rights perspectives, this is good. However, when concerns were raised, they "were usually not tied directly to either national laws or international norms nor to Facebook's values such as voice."[38]

---

[31] *Id.*

[32] KETTEMANN & SCHULZ, *supra* note 1, at 18.

[33] KETTEMANN & SCHULZ, *supra* note 1, at 18.

[34] KETTEMANN & SCHULZ, *supra* note 1, at 20 (quoting Monika Bickert).

[35] KETTEMANN & SCHULZ, *supra* note 1, at 19 ("Privacy and dignity are constitutional values that are explicitly protected in all liberal democracies and by the International Bill of Rights and regional human rights conventions. The Community Standards do not explicitly refer to these documents…[t]he same is true for the described method of 'weighing the public interest value against the risk of harm'.").

[36] *Product Policy Forum Minutes*, FACEBOOK (Nov. 15, 2018), https://about.fb.com/news/2018/11/content-standards-forum-minutes.

[37] *Stakeholder Engagement: How Does Stakeholder Engagement Help Us Develop Our Community Standards?*, FACEBOOK, https://www.facebook.com/communitystandards/stakeholder_engagement (last visited July 21, 2020).

[38] KETTEMANN & SCHULZ, *supra* note 1, at 26.

Facebook also tends to escalate "normative change processes" to senior leadership (typically Mark Zuckerberg, Sheryl Sandberg, Joel Kaplan and Nick Clegg)[39] but how it does this is completely opaque. According to Kettemann and Schultz, "[t]he role of integrating leadership feedback here seems to rest, as has been described in a number of journalistic pieces, with Monika Bickert, the VP for Global Policy Management, who would, if needed, 'take it [i.e. the issue] to Mark [Zuckerberg].' We could not shed more light on this. Our methods of observation unfortunately had their limits."[40]

Facebook states that the balancing test helps to reconcile conflicts among values, but in reality, the test is applied secretly and without reference to any structured, public, and transparent framework. Looking to IHRL as guidance is the way for Facebook to address these problems. Facebook's current processes, values, and Community Standards are drafted and modified solely by Facebook. By contrast, IHRL was developed independent of Facebook or any other social media company's influence. The ICCPR has 173 state parties[41] and the Genocide Convention has 147 parties. Many of their provisions can be said to reflect customary international law. The ICCPR maps freedom of expression and appropriate restrictions thereof, which appears to overlap with Facebook's content moderation efforts. Importantly, the Oversight Board can cite and rely on IHRL as a source of law, rather than a quasi-legal construct of the Community Standards and Facebook's internal debates.

Facebook itself has said that "[w]e look for guidance in documents like Article 19 of the ICCPR, which set standards for when it's appropriate to place restrictions on freedom of expression."[42] The ICCPR maintains that everyone has the right to freedom of expression and that restrictions on this right are only allowed when they are "provided by law

---

[39] *See* Elizabeth Dwoskin & Nitasha Tiku, *Facebook Employees Said They Were 'Caught in an Abusive Relationship' with Trump as Internal Debates Raged*, WASH. POST (June 5, 2020), https://www.washingtonpost.com/technology/2020/06/05/facebook-zuckerberg-trump/ [https://perma.cc/H94L-AS9V] ("In addition to diversity head Williams, the team that made the decision included Zuckerberg; Sandberg; Joel Kaplan, the vice president for U.S. public policy; and Nick Clegg, the vice president of global affairs and communications; as well as the head of human resources and the general counsel."); Sheera Frankel, *Delay, Deny and Deflect: How Facebook's Leaders Fought Through Crisis*, N.Y. TIMES (Nov. 14, 2018), https://www.nytimes.com/2018/11/14/technology/facebook-data-russia-election-racism.html [https://perma.cc/9Y85-VRUN] ("Some at Facebook viewed Mr. Trump's 2015 attack on Muslims as an opportunity to finally take a stand against the hate speech coursing through its platform. But Ms. Sandberg, who was edging back to work after the death of her husband several months earlier, delegated the matter to Mr. Schrage and Monika Bickert, a former prosecutor whom Ms. Sandberg had recruited as the company's head of global policy management. Ms. Sandberg also turned to the Washington office — particularly to Mr. Kaplan, said people who participated in or were briefed on the discussions.").

[40] KETTEMANN & SCHULZ, *supra* note 1, at 29.

[41] Office of United Nations High Commissioner for Human Rights, *Status of Ratification: Interactive Dashboard*, https://indicators.ohchr.org (last visited July 21, 2020).

[42] Richard Allan, *Hard Questions: Where Do We Draw the Line on Free Expression?*, FACEBOOK (Aug. 9, 2018), https://about.fb.com/news/2018/08/hard-questions-free-expression.

and are necessary for: (a) the respect of the rights or reputations of others; (b) for the protection of national security or of the public order, or of public health or morals."[43] However, Facebook has adopted a conclusory interpretation of Article 19, noting, "[t]he core concept here is whether a particular restriction of speech is necessary to prevent harm. Short of that, the ICCPR holds that speech should be allowed. This is the same test we use to draw the line on Facebook."[44] By collapsing the ICCPR Article 19(3) tests of legality, legitimacy, and necessity and proportionality into an undefined "risk of harm" test, Facebook has failed to adopt a real test at all, and has instead maintained the current regime of ad-hoc content decision making. In that same post, Facebook goes on to apply a harm-based analysis independent of what the ICCPR says.[45] Consistent with Kettemann and Schulz's findings, Facebook continues to wield undefined discretion. Harm is how Facebook itself defines it.

B. Applying IHRL to Content Decisions Gives Social Media Companies Legitimacy and Provides One Universal Standard

As we can see, Facebook's current interpretation of IHRL is vague. IHRL is referenced as a source of 'guidance,' while moderation in practice continues to be *ad hoc* and non-transparent. This section argues that IHRL, if correctly applied, can provide legitimacy to Facebook's content moderation decisions. There is a strong argument to be made that Facebook and its Oversight Board have willingly bound themselves to IHRL. Facebook has stated[46] that it will comply with the UN Guiding Principles on Business and Human Rights (the "UNGPs").[47] Facebook's agreement to comply with the UNGPs essentially prevents Facebook from claiming that IHRL applies only to states. Guiding Principle (GP) 11 states that "[b]usiness enterprises should respect human rights."[48] GP 12 interprets "respect human rights" to mean "at a minimum, as those expressed in the International Bill of Human Rights and the principles concerning fundamental rights set out in the International Labour Organization's Declaration on Fundamental Principles and Rights at Work."[49] The commentary to GP 12 states that the International Bill of Human Rights consists of "the Universal Declaration of Human Rights and the main instruments through which it has been codified: The International

[43] ICCPR, *supra* note 19, art. 19.
[44] Allan, *supra* note 42, at para. 5.
[45] Allan, *supra* note 42, at para. 9.
[46] Alex Warofka, *Human Rights Impact of Facebook in Myanmar*, FACEBOOK (Nov. 5, 2018), https://about.fb.com/news/2018/11/myanmar-hria.
[47] Office of the United Nations High Commissioner for Human Rights, *Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, UN Doc HR/PUB/11/04 (2011), https://www.ohchr.org/documents/publications/guidingprinciplesbusinesshr_en.pdf.
[48] *Id.* at 13.
[49] *Id.*

Covenant on Civil and Political Rights and the International Covenant on Economic, Social and Cultural Rights."[50] GP 31 deals with the "effectiveness criteria for non-judicial grievance mechanisms," which is exactly what the Oversight Board and Facebook's appeals processes for content moderation are.[51] GP 31 outlines eight criteria "to ensure their effectiveness":

a) **Legitimate**: enabling trust from the stakeholder groups for whose use they are intended, and being accountable for the fair conduct of grievance processes;
b) **Accessible**: being known to all stakeholder groups for whose use they are intended, and providing adequate assistance for those who may face particular barriers to access;
c) **Predictable**: providing a clear and known procedure with an indicative time frame for each stage, and clarity on the types of process and outcome available and means of monitoring implementation;
d) **Equitable**: seeking to ensure that aggrieved parties have reasonable access to sources of information, advice and expertise necessary to engage in a grievance process on fair, informed and respectful terms;
e) **Transparent**: keeping parties to a grievance informed about its progress, and providing sufficient information about the mechanism's performance to build confidence in its effectiveness and meet any public interest at stake;
f) **Rights-compatible**: ensuring that outcomes and remedies accord with internationally recognized human rights;
g) **A source of continuous learning**: drawing on relevant measures to identify lessons for improving the mechanism and preventing future grievances and harms;

Operational-Level Mechanisms should also be:

h) **Based on engagement and dialogue**: consulting the stakeholder groups for whose use they are intended on their design and performance and focusing on dialogue as the means to address and resolve grievances.[52]

In applying GP 31 to Facebook's own content moderation tests, one sees how inadequate Facebook's tests are. The manner in which Facebook currently makes content decisions does not enable trust; these con-

[50] *Id.* at 14.
[51] *Id.* at 33.
[52] *Id.* at 35.

tent moderation processes are famously opaque.[53] Content decisions are not accessible or predictable.[54] In 2017, journalists forced Facebook to admit to making mistakes in decisions on over half of the cases of hate speech that the journalists identified.[55] One might say that the Community Standards have increased equity, but the grievance process is still opaque. It took a year after the leak of its internal guidance for content moderators before Facebook published information on how content moderation decisions are made in practice. At the same time, Facebook launched, for the first time, an appeals process for wrongfully removed of content.[56] *Ad hoc* decision-making not only obscures how outcomes and remedies accord with IHRL but also impedes continuous learning. Only recently has Facebook increased engagement and dialogue with stakeholder groups, including the Oversight Board itself.[57]

The issues flagged by GP 31 can be addressed by respecting and implementing the IHRL in a transparent, predictable, and structured manner. With a clearly structured framework for the Oversight Board to apply, its members would be inclined to use the ICCPR for decision making. This in turn would give Facebook, the Board, and a future Social Media Council both independence and legitimacy for content decisions. Facebook could start with the key multilateral treaty on freedom of expression, the ICCPR.[58] The key challenge here concerns how to apply the ICCPR, which was written for states, to powerful non-state actors.

This essay does not argue that Facebook and the Oversight Board should adopt *all* IHRL sources. Instead, it recommends that both Facebook and the Oversight Board tread carefully by first starting with implementation of ICCPR Article 19. Some IHRL sources, such as Article 20 of the ICCPR, need significant clarification by scholars, practitioners, and other observers before they can be applied to Facebook content moderation rules.[59] The Board and commentators will have to do work to make clear what Article 20 terms like "incitement" and "advocacy to hatred" mean as applied to Facebook. Where other sources of IHRL con-

---

[53] Sarah Roberts, *Digital Detritus: 'Error' and the Logic of Opacity in Social Media Content Moderation*, 23:3 FIRST MONDAY (2018), https://firstmonday.org/ojs/index.php/fm/article/view/8283/6649.

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

[55] Madeleine Varner, Ariana Tobin, Julia Angwin & Jeff Larson, *What Does Facebook Consider Hate Speech?*, PROPUBLICA (Dec. 28, 2017) https://projects.propublica.org/graphics/facebook-hate [https://perma.cc/93UP-D58X].

[56] Julia Carrie Wong & Olivia Solon, *Facebook Releases Content Moderation Guidelines–Rules Long Kept Secret*, THE GUARDIAN, (Apr. 24, 2018), https://www.theguardian.com/technology/2018/apr/24/facebook-releases-content-moderation-guidelines-secret-rules [https://perma.cc/2GJV-7CUE].

[57] Warofka, *supra* note 46, paras. 6, 34.

[58] International Covenant on Civil and Political Rights, Dec. 16, 1966, T.I.A.S. 92-908, 999 U.N.T.S. 171.

[59] *Id.*, art. 20.

flict with the ICCPR Article 19, the Article 19 of the ICCPR should control.

C. The ICCPR Is the Scaffolding, But IHRL Still Needs to be Concretely Adapted to Social Media Companies

Facebook has agreed to "respect human rights" under the UNGPs and has stated that it will look to the ICCPR for guidance. The next step is to determine how the ICCPR can help rather than hinder social media companies in making content decisions. The proposed framework is as follows. Facebook will no longer be in the business of "making up rules." Instead, whenever Facebook decides to draft or amend its Community Standards and make content reviews and decisions, its leadership and the Oversight Board should adopt the following procedure.

1. Adapt ICCPR Article 20 to social media companies in a way that is comprehensible. Being part of IHRL, Article 20 cannot be ignored. At present, however, Article 20(2) and the Rabat definitions[60] interpreting Article 20(2) are almost incoherent. Significant interpretive guidance on Article 20 must be provided by IHRL institutions, commentators, and the Oversight Board itself.
2. Apply the adapted ICCPR Article 19(3) test–to be examined in Part III–to see whether prohibition of the content in question may be permissible. Some guiding questions include:
    a. Did Facebook draft and issue the content prohibition in accordance with our adapted legality test?
    b. Does Facebook have a legitimate interest(s) in the content prohibition?
    c. In light of the adapted Rabat 6-factor test, is Facebook's content prohibition necessary and proportionate to the legitimate interest Facebook has asserted?

III. Whose Legality, Legitimacy, Necessity and Proportionality?

In this section, I describe existing problems with Article 20 of the ICCPR, which complicates its adaptability to social media companies. By comparison, Article 19 of the ICCPR is much easier to adapt. I articulate how the three prongs of Article 19–(i) legality, (ii) legitimacy, and (iii) necessity and proportionality–can be workably adapted to content moderation by Facebook and other social media companies.

[60] As will be discussed later in this piece, the Rabat Plan of Action makes attempts to define the terms in Article 20 terms. However, I believe that Rabat fails in this regard. *See infra* note 66.

A. Article 20: Not the Best Written Article of the ICCPR

Under the ICCPR, Article 19 is not the only article that governs freedom of expression. Per David Kaye, the UN Special Rapporteur for Freedom of Expression, "[u]nder article 20(2) of the Covenant, States parties are obligated to prohibit by law 'any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence.' States are not obligated to criminalize such kinds of expression."[61] ICCPR Article 20 also requires that "[a]ny propaganda for war shall be prohibited by law."[62] Adapted to social media companies, Article 20 sets the standard by which social media companies are *required* to restrict certain kinds of content. Of course, since social media companies do not have law making powers, their adapted Article 20(2) obligations become that of prohibiting such content via their Terms of Service (e.g. Facebook's Community Standards).[63]

ICCPR Article 20 suffers from vagueness. Neither "propaganda for war" nor "constitutes incitement" is defined. In response to this issue, ARTICLE 19 drafted the Camden Principles, which proposes definitions for some of the terms in Article 20, namely "hatred," "hostility," "advocacy," and "incitement."[64] Similarly, the UN Office of the High Commissioner for Human Rights (OHCHR) has also completed a four-year initiative to clarify Article 20.[65] The Rabat Plan of Action ("Rabat") is the outcome of that initiative. While Rabat provides sample definitions of Article 20 terms for states to implement–which were adapted from the Camden Principles–it unfortunately muddies them by combining "incitement" and "hatred" while at the same time retaining separate definitions for each.[66] How muddied are the Rabat definitions of Article 20 terms? Let's see:

---

[61] Kaye, *supra* note 22, at para. 8.

[62] ICCPR, *supra* note 19, art. 20(1).

[63] Kaye, *supra* note 22, at para. 10 ("[a] critical point is that the individual whose expression is to be prohibited under article 20 (2) of the Covenant is the advocate whose advocacy constitutes incitement. A person who is not advocating hatred that constitutes incitement to discrimination, hostility or violence, for example, a person advocating a minority or even offensive interpretation of a religious tenet or historical event, or a person sharing examples of hatred and incitement to report on or raise awareness of the issue, is not to be silenced under article 20 (or any other provision of human rights law).").

[64] ARTICLE 19, *The Camden Principles on Freedom of Expression and Equality* (Apr. 2009) https://www.article19.org/data/files/pdfs/standards/the-camden-principles-on-freedom-of-expression-and-equality.pdf.

[65] ARTICLE 19, *ARTICLE 19 welcomes the Rabat Plan of Action on Prohibition of Incitement and Calls for its Full Implementation* (Nov. 16, 2012), https://www.article19.org/resources/article-19-welcomes-rabat-plan-action-prohibition-incitement-calls-full-implementation.

[66] *See* U. N., *Rabat Plan of Action on the Prohibition of Advocacy of National, Racial or Religious Hatred that Constitutes Incitement to Discrimination, Hostility or Violence*, A/HRC/22/17/Add.4 (Oct. 5, 2012) https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf [hereinafter Rabat]; *see also* Rabat, 10 n. 5, (all other definitions); Rabat, 9 n. 21 (Rabat prefers that states "consider including robust definitions of key terms such as hatred, discrimination, violence, hostility, among others").

- Rabat utilizes "incitement to hatred" as a catch-all term for "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[67] Yet incitement also "refers to statements about national, racial or religious groups which create an imminent risk of discrimination, hostility or violence against persons belonging to those groups."[68]
- "Hatred" and "hostility" somehow have the same definition. They "refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group."[69] This is confusing because "incitement to hatred" includes not only hostility but also "discrimination" and "violence."
- Advocacy "is to be understood as requiring an intention to promote hatred publicly towards the target group."[70] But "hatred" is used in advocacy's singular definition as well as the catch-all definition of "incitement to hatred," which includes incitement to discrimination, hostility, or violence. In the overall definition, "hatred" is also used, but separate from "advocacy," in the language "any advocacy of national, racial, or religious hatred."[71]

To say that this is confusing is an understatement. Rabat's Article 20 definitions are riddled with tautologies.

Under the most charitable reading, the Rabat Plan seems to prohibit intense and irrational public speech that intends to incite discrimination, hostility, or violence based on nationality, race, or religion. Thus, states would be required to prohibit content fitting this reading. "Violence" perhaps has a straightforward definition of physical harm against people. Rabat itself and human rights INGOs have left states to interpret the IHRL definitions of "discrimination" and "hostility" when implementing domestic laws for the ICCPR. Facebook and the Oversight Board will likely have similar definitional duties. Facebook and the Oversight Board will also have to define "propaganda for war" in Article 20(1)–perhaps by drawing on IHRL sources such as the Geneva Conventions.

In summary, when Facebook and the Oversight Board are deciding what speech they are required to prohibit, they must look to ICCPR Article 20(1) and 20(2)–perhaps employing the tests I have proposed above. Due to the confusion surrounding Article 20 and Rabat, however, this

[67] *Id.* at 6 n. 1 (defining "incitement to hatred").
[68] *Id.* at 10 n. 5 (defining "incitement").
[69] *Id.* at 10 n. 5.
[70] *Id.*
[71] *Id.* at 6 n. 1.

will not be the focus of this Essay. Instead, I shall focus on a clearer test, which is found in Article 19 of the ICCPR.[72]

B. Article 19: When Content May Be Restricted

Much of the credit for establishing how IHRL may be applied to social media companies goes to the text of the ICCPR itself and UNHRC's General Comment No. 34. David Kaye has taken both the ICCPR text and General Comment No. 34 and provided further analysis. In his October 2019 submission, Kaye writes that the ICCPR has long been the key IHRL source of law, noting that ICCPR Article 19(1) "protects the right to hold opinions without interference" and Article 19(2) "guarantees the right to freedom of expression, that is, the right to seek, receive and impart information and ideas of all kinds, regardless of frontiers, through any media."[73] In that submission, Kaye also discusses Article 19(3), which limits expression under Article 19(2) "only where provided by law and necessary to respect the rights or reputations of others or protect national security, public order, public health or morals."[74] Noting that these are "narrowly defined exceptions . . . the burden falls on the authority restricting speech to justify the restriction, not on the speakers to demonstrate that they have the right to such speech."[75]

When Facebook is deciding how to craft its Community Standards and algorithms on content it *decides* to restrict (rather than *required* by IHRL to prohibit), it should rely on the three-prong test articulated in ICCPR Article 19(3). When the Oversight Board is *reviewing* Facebook's decision on a piece of content, the Oversight Board should assess whether Facebook has acted, both in its actions and the Community Standards, in accordance with ICCPR Article 19(3). Besides the ICCPR, General Comment No. 34 also offers an authoritative interpretation of the (i) legality, (ii) legitimacy, and (ii) necessity and proportionality prongs of Article 19(3).[76]

Legality: "The restriction must be provided by laws that are precise, public and transparent; it must avoid providing authorities with un-

---

[72] Note that Rabat has stated that the "higher threshold" in Article 20 must "take into account the provisions of article 19… the three-part test (legality, proportionality and necessity) for restrictions also applies to cases involving incitement to hatred, in that such restrictions must be provided by law, be narrowly defined to serve a legitimate interest, and be necessary in a democratic society to protect that interest". *Id.* at 9 para. 18. As a practical matter and as will be discussed later in this essay, the necessity and proportionality prong of Article 19 should limit Article 20 to *require prohibition* of speech only for the most intense and irrational public speech that intends to incite discrimination, hostility, or violence based on nationality, race, or religion, in accordance with the Rabat 6-factor test–lesser speech should be subject to sanction less than prohibition.

[73] Kaye, *supra* note 22, at para. 5.

[74] Kaye, *supra* note 22, at para. 6.

[75] Kaye, *supra* note 22, at para. 6.

[76] UNHRC General Comment No. 34, para. 6, CCPR/C/GC/34 (Sept. 12, 2011), https://www2.ohchr.org/english/bodies/hrc/docs/GC34.pdf.

bounded discretion, and appropriate notice must be given to those whose speech is being regulated. Rules should be subject to public comment and regular legislative or administrative processes. Procedural safeguards, especially those guaranteed by independent courts or tribunals, should protect rights."[77]

Legitimacy: "The restriction should be justified to protect one or more of the interests specified in article 19 (3) of the Covenant, that is, to respect the rights or reputations of others or to protect national security, public order, public health or morals."[78]

Necessity and Proportionality: "The restriction must be demonstrated by the State as necessary to protect a legitimate interest and to be the least restrictive means to achieve the purported aim. The Human Rights Committee has referred to these conditions as 'strict tests', according to which restrictions 'must be applied only for those purposes for which they were prescribed and must be directly related to the specific need on which they are predicated.'"[79]

IV. Adapting the ICCPR to Social Media Companies (Non-States)

Despite being a good starting point, David Kaye's analysis fails to take note of the fact that since the ICCPR is a multilateral treaty adopted by *states* (countries), there is language in Article 19(3) that does not make sense in the context of *social media companies*. I argue that the language of Article 19(3) needs to be adapted to fit the needs of social media companies. This section breaks down how social media companies can apply each of the three tests from Article 19(3)–(i) legality, (ii) legitimacy, and (iii) necessity and proportionality–to their content moderation decisions.

A. Adapting Legality to Social Media Companies

Unlike countries, social media companies cannot pass laws because they are not states imbued with a legislative function by the consent of the governed. Nor can social media companies guarantee the protection of rights using procedural safeguards by independent courts or tribunals, since they are not states imbued with a judicial function. How can a "legality" prong be applied to social media companies, who have no powers of law? I suggest that IHRL provides the answer. 173 member states have consented to the ICCPR. The ICCPR's "legality" prong imposes re-

[77] Kaye, *supra* note 22, at para. 6(a).
[78] Kaye, *supra* note 22, at para. 6(b)
[79] Kaye, *supra* note 22, at para. 6(c). Put differently, there can be valid restrictions on freedom of expression (i) if the restriction is a valid law subject to public review and judicial review, (ii) if the restriction is justified by one of the enumerated legitimate interests, (iii) if the restriction is the least restrictive means to protect a justifiable legitimate interest, and (iv) if the restriction is directly related to that interest.

quirements on states to guarantee the protection of rights through procedural safeguards, including independent review. Recall that Facebook has already consented to the UNGPs as well ICCPR Article 19. Imposing an IHRL framework on social media companies' content moderation rules and independent review of those rules through the Oversight Board is the answer. The question then becomes how to adapt the "legality" prong to social media companies.

As applied to Facebook, its Community Standards are currently not in compliance with this adapted legality prong. While Facebook's standards are "public," they are neither "precise" nor "transparent." As INGO ARTICLE 19 notes, many terms are not defined, and the Community Standards are changed often without any notice or rationale given.[80] Facebook currently has "unbounded discretion"[81] in coming up with and implementing the Community Standards. Proposed edits to the Community Standards are neither "subject to public comment" nor "regular...administrative processes."[82] The pending Oversight Board could provide the relevant "procedural safeguards, especially those guaranteed by independent courts or tribunals" to "protect rights," but this remains to be seen.[83]

The legality prong is applicable to the algorithms Facebook uses to determine a post's virality, as well as takedowns of posts that presumptively contain "hate speech" or incitement of violence. I have suggested to the Oversight Board staff that independent audits of Facebook's algorithms and human content moderators should fall under the Oversight Board's jurisdiction. The "procedural safeguards" of the legality prong may require "human-in-the-loop" machine learning, whereby humans are involved in training and testing algorithms.[84]

B. Adapting Legitimacy to Social Media Companies

The legitimate interests of states are not equivalent to the legitimate interests of social media companies. As applied to social media companies, legitimate interests generally fall within a spectrum. On the one end are national security interests, which I believe social media companies cannot invoke. They cannot claim national security as a legitimate inter-

---

[80] ARTICLE 19, *Facebook Community Standards: Analysis Against International Standards on Freedom of Expression* (Jul. 30, 2018) https://www.article19.org/resources/facebook-community-standards-analysis-against-international-standards-on-freedom-of-expression. Note that Facebook has made effort to define some terms since this report, showing the importance of independent and public analysis.

[81] Kaye, *supra* note 22, at para. 6(a).

[82] Kaye, *supra* note 22, at para. 6(a).

[83] Kaye, *supra* note 22, at para. 6(a).

[84] *See* Kaye, *supra* note 22, at para. 58(d) ("Ensure that any enforcement of hate speech rules involves an evaluation of context and the harm that the content imposes on users and the public, including by ensuring that any use of automation or artificial intelligence tools involve human-in-the-loop").

est to restrict speech or say, give speech (e.g. user data) to governments. On the other end are public health interests, which social medial companies can invoke and which are the easiest to defend. All other interests fall somewhere in between.

1. National Security

National security, with one important exception, is not a "legitimate interest" for social media companies given that they have no "national" security interests to protect. General Comment No. 34 notes that states must take care to ensure that laws limiting expression related to "official secrets," "sedition laws or otherwise," and "prosecut[ing] journalists, researchers, environmental activities, human rights defenders, or others" be extremely narrowly tailored in accordance with Article 19(3).[85] At any rate, none of those examples are within social media companies' powers.[86] If social media companies were to deploy similar restrictions (e.g. sedition communications against states), the Oversight Board or Social Media Council should find these restrictions illegitimate.[87]

The important exception is that social media companies may use the "national security" interest as a rationale against following state-asserted national security claims (e.g. the Chinese government stating that they need TikTok's user data for national security purposes). TikTok, in this instance, may validly claim that since it is a social media company, and not a nation, it cannot assert a national security rationale as a reason to restrict or provide speech.

2. Rights and Reputations of Others

General Comment No. 34, paragraph 28, advances the "rights and reputations of others" as a legitimate interest through citation to HRC decisions:

> The term 'rights' includes **human rights** as recognized in the Covenant and more generally in international human rights law. For example, it may be legitimate to restrict freedom of expression in order to protect the right to vote under article 25, as well as rights under article 17 . . . Such restrictions must be constructed with care: while it may be permissible to protect vot-

[85] General Comment No. 34, *supra* note 76, at para. 30.

[86] *See* Kaye, *supra* note 22, at para. 30.

[87] Some have questioned whether social media companies should be applying national security interest rules to protecting the security *of states* (e.g. preventing terrorists from using Facebook to coordinate bombings). The use of these interests to restrict content, however, may open the door for sovereigns to accuse social media companies of defining what constitutes a valid "national security" interest, which has traditionally been the role of states and applicable tribunals, not social media companies. States may then argue that social media companies should instead act in compliance with "national security" domestic laws, regardless of whether or not these laws are compliant with IHRL. In light of this danger, it is better for social media companies to rely instead on the legitimate interest of "public order"–discussed below–in order to protect their own *users* from violence.

> ers from forms of expression that constitute intimidation or coercion, such restrictions must not impede political debate, including, for example, calls for the boycotting of a non-compulsory vote."[88]

Facebook appears to manifest this voting interest through a Newsroom post linked to its Manipulated Media,[89] Violence and Incitement,[90] and Coordinating Harm and Publicizing Crime Community Standards.

Facebook also has Community Standards governing intellectual property rights[91] and privacy rights.[92] Article 15.1(c) of the International Covenant on Economic, Social and Cultural Rights (ICESCR) recognizes the rights of everyone "[t]o benefit from the protection of the moral and material interests resulting from any scientific, literary or artistic production of which he is the author."[93] UNGP 12, which Facebook has consented to following, lists ICESCR in its commentary as one of the sources of IHRL that business enterprises must follow to respect human rights.[94] Therefore, the protection of intellectual property rights is a valid application of the "rights and reputations of others" as a legitimate interest under the proposed ICCPR Article 19 test.[95]

It is also likely that the rights and reputations of others interest applies to the "public figure" concept in law. General Comment No. 34 notes that the ICCPR places high value on protecting uninhibited expression in public debate concerning public figures in the political domain and public institutions: "All public figures, including those exercising the highest political authority such as heads of state and government, are legitimately subject to criticism and political opposition."[96] As will be shown in the adapted Rabat test I propose later in this essay, some restrictions on freedom of expression towards public figures may be impermissible given their public speaker status, but other restrictions may

---

[88] General Comment No. 34, *supra* note 76, at para. 28 (emphasis added).

[89] *See* Kevin Martin & Samidh Chakrabarti, *Helping to Protect the 2020 US Census*, FACEBOOK (Dec. 19, 2019), https://about.fb.com/news/2019/12/helping-protect-the-us-census.

[90] *Violence and Incitement: Policy Rationale*, FACEBOOK (June 2020), https://www.facebook.com/communitystandards/recentupdates/all_updates ("Any content containing statements of intent, calls for action, or advocating for high or mid-severity violence due to voting, voter registration, or the outcome of an election. Misrepresentation of the dates, locations, and times, and methods for voting or voter registration. . . . Misrepresentation of who can vote, qualifications for voting, whether a vote will be counted, and what information and/or materials must be provided in order to vote. . . . Other misrepresentations related to voting in an official election may be subject to false news standards, as referenced in section 18.").

[91] *Community Standards*, FACEBOOK, https://www.facebook.com/communitystandards/intellectual_property (last visited July 21, 2020).

[92] *Privacy Violation and Image Privacy Rights*, FACEBOOK, https://www.facebook.com/communitystandards/privacy_violations_image_rights (last visited July 21, 2020).

[93] U.N. International Covenant on Economic, Social and Cultural Rights, art. 15, Dec. 16, 1966, 993 U.N.T.S. 3.

[94] *Guiding Principles on Business and Human Rights*, *supra* note 47.

[95] ICCPR, *supra* note 19, art. 19.

[96] General Comment No. 34, *supra* note 76, at para. 38.

be permissible if the public speaker is using their position to spread content that runs afoul of one of the legitimate interests in Article 19.

3. Public Order

General Comment No. 34, paragraph 31, notes that "it may, for instance, be permissible in certain circumstances to regulate speech-making in a particular public place," mentioning "[c]ontempt of court proceedings" as an example97. Thus, social media companies may cite "public order" interests to prevent incitement of violence, disturbance of peace, or other criminal activities that might arise on their platforms.98

Social media companies may also have a commercial interest in maintaining public order on their platforms so as to not disrupt the user experience. Such disruptions may alienate users or cause them to leave. For example, imagine a scenario where an app like Tinder, which has an interest in maintaining a good dating user experience for its users, gets overrun by users making trolling or offensive comments on their Tinder profiles. Tinder's users start leaving the platform for other dating apps due to the hostile user experience. Tinder may reasonably prohibit such content in order to maintain its central dating user experience. Thus narrowly tailored restrictions on user safety,99 hate speech,100 pornography, and cyberbullying are appropriate under the public order interest, and Facebook has Community Standards for these categories.101 As

---

[97] General Comment No. 34, *supra* note 76, at para. 31.

[98] *See, e.g.*, *Violence and Criminal Behavior*, FACEBOOK, https://www.facebook.com/communitystandards/violence_criminal_behavior (last visited July 21, 2020). However, the legitimacy prong in tandem with the legality prong reveals several issues with how Facebook has drafted this section. For example, for "dangerous individuals and organizations", Facebook provides no precision or process or independent review of how it decides that an organization is "proclaim[ing] a violent mission or [is] engaged in violence to have a presence on Facebook."

[99] *Safety*, FACEBOOK, https://www.facebook.com/communitystandards/safety (last visited July 21, 2020).

[100] Given the use of hate speech to incite violence in Myanmar and other countries, the legitimate interest of public order should apply. In terms of sources of IHRL, the UN Strategy and Plan of Action on Hate Speech is controlling. UN, *Strategy and Plan of Action on Hate Speech*, 2 (May 2019), https://www.un.org/en/genocideprevention/documents/UN%20Strategy%20and%20Plan%20of%20Action%20on%20Hate%20Speech%2018%20June%20SYNOPSIS.pdf. (providing a definition for "hate speech" as "any kind of communication in speech, writing or behaviour, that attacks or uses pejorative or discriminatory language with reference to a person or a group on the basis of who they are, in other words, based on their religion, ethnicity, nationality, race, colour, descent, gender or other identity factor.") The document also notes that content containing this alone is insufficient, but requires also "incitement to discrimination, hostility and violence", an apparent reference to ICCPR Article 20(2), (which again leaves the outstanding issue of needing to clear up exactly what Article 20(2) means, while also noting that any Article 20 analysis must be in conformance with Article 19's requirements). *Id.* at 1.

[101] *See Adult Nudity and Sexual Activity*, FACEBOOK, https://www.facebook.com/communitystandards/recentupdates/adult_nudity_sexual_activity (last visited July 21, 2020); *Bullying and Harassment*, FACEBOOK, https://www.facebook.com/communitystandards/recentupdates/bullying (last visited July 21, 2020).

Kettemann and Schulz note, the fact that "we did not notice any reference to economic interests, much less controlling economic interests in the process, does of course not mean that they are irrelevant."102 It is important to emphasize that since Facebook covers a much broader set of human expression as part of its user experience–as compared to Tinder– this suggests that more kinds of expression on Facebook are permissible and in line with user experience. This means that restrictions of expressions on Facebook must be narrowly tailored.

The scope of the Oversight Board's powers is in dispute as of the writing of this piece. Some commentators worry that the Oversight Board will only have "thumbs up, thumbs down" powers over one-off content moderation decisions rather than the power to review the machine learning, technical, and operational elements involved in content moderation which would have large scale impact.103 Such efforts not only require power of oversight over algorithms and automated content moderation, but also human-coordinated propaganda that are often much more effective in spreading disinformation, misinformation, and malinformation.104 Public order interests should be interpreted to grant the Board broad powers to make these important decisions.

Currently, the Oversight Board's charter and bylaws are vague on the scope of the Board's powers. Their powers are also limited by procedure. The Board currently has no sua sponte powers105 and must wait for a case to be brought before the Board for review, either through an appeals process or by the urging of Facebook itself.106 I suggest that the Board be granted sua sponte powers, either by its own Marbury107 fiat in a case decision, or in updated bylaws that the Oversight Board intends to release later this year. Such sua sponte powers should grant the Board authority to initiate its own cases based on its observations of trends on the Facebook platform and beyond.

---

[102] KETTEMANN & SCHULZ, *supra* note 1, at 31.

[103] Matthew Ingram, *Alex Stamos Talks About Facebook's Oversight Board*, GALLERY BY CJR https://galley.cjr.org/public/conversations/-M74eLMfvkdKpIPjRfo4 (last visited Jul. 21, 2020).

[104] UNESCO, *Journalism, 'Fake News' and Disinformation: A Handbook for Journalism Education and Training* (2018), https://en.unesco.org/fightfakenews.

[105] *See* Legal Information Institute (LII), *Sua sponte*, CORNELL L. SCH., https://www.law.cornell.edu/wex/sua_sponte ("Latin for 'of one's own accord; voluntarily.' Used to indicate that a court has taken notice of an issue on its own motion without prompting or suggestion from either party.").

[106] *See Oversight Board Charter,* art. 2, § 1, FACEBOOK ("In instances where people disagree with the outcome of Facebook's decision and have exhausted appeals, a request for review can be submitted to the board by either the original poster of the content or a person who previously submitted the content to Facebook for review. Separately, Facebook can submit requests for review, including additional questions related to the treatment of content beyond whether the content should be allowed or removed completely.").

[107] *See Marbury v. Madison*, 5 U.S. 137, 177 (1803) ("It is emphatically the province and duty of the Judicial Department to say what the law is. Those who apply the rule to particular cases must, of necessity, expound and interpret that rule.").

4. Public Health

In contrast to "national security", public health may be the interest most readily applicable to social media companies who have an interest in battling misinformation on their platforms that may affect the health of their users. Facebook's current efforts to combat coronavirus misinformation are representative of this interest in action.[108] For example, online illegal wildlife trade is growing in recent years, especially on Facebook.[109] It may be that the next COVID-19 pandemic emerges from the illegal trading of wildlife online. Facebook currently relies on NGOs such as the World Wildlife Fund (WWF) to give them notice of such trades. However, having a structured IHRL public health framework would be superior to Facebook's current *ad hoc* approach to public health issues. The Oversight Board could find that the WHO Constitution[110] and the International Health Regulations (IHR)[111] must be adapted to Facebook's content moderation regulations and procedures. This would be in line with Facebook's commitment to following the UNGPs. Article 6 of the IHR may require Facebook to give timely notice to the WHO "of all events which may constitute a public health emergency of international concern."[112] This would create an obligation on Facebook to *proactively* monitor and report future public health emergencies like COVID-19.

Such a requirement would not be too onerous as Facebook already does this for child pornography. Child pornography and illegal wildlife trade are amendable to computer vision solutions, such as PhotoDNA, which Microsoft developed and has donated to the National Center for Missing & Exploited Children (NCMEC). Currently, Facebook uses to this technology to detect child pornography.[113]

---

[108] Kang-Xang Jin, Head of Health, *Keeping People Safe and Informed About the Coronavirus*, FACEBOOK (June 24, 2020), https://about.fb.com/news/2020/06/coronavirus [https://perma.cc/6ZM3-RJBH]; Virginia Allen, *What Does Facebook's New Oversight Board Mean for Conservative Posts?*, THE DAILY SIGNAL (May 14, 2020), https://www.dailysignal.com/2020/05/14/what-does-facebooks-new-oversight-board-mean-for-conservative-posts [https://perma.cc/9MV6-DU7S].

[109] *See, e.g.*, Thu Thu Aung, *Facebook Purges Ads for Illegal Wildlife in Southeast Asia as Online Trade Surges*, REUTERS (Aug. 6, 2020), https://www.reuters.com/article/us-myanmar-wildlife/facebook-purges-ads-for-illegal-wildlife-in-southeast-asia-as-online-trade-surges-idUSKCN2520C3 [https://perma.cc/83P7-7UMA
] ("in the five months through May 2020, a report seen by Reuters showed World Wildlife Fund researchers had counted 2,143 wild animals from 94 species for sale on Facebook from Myanmar alone.").

[110] Constitution of the World Health Organization, WORLD HEALTH ORGANIZATION (WHO) (July 22, 1948), https://apps.who.int/gb/bd/PDF/bd47/EN/constitution-en.pdf?ua=1. The WHO has 194 member states as of this writing.

[111] International Health Regulations (IHR), WORLD HEALTH ORGANIZATION (WHO) (2d ed., 2005), https://apps.who.int/iris/bitstream/handle/10665/43883/9789241580410_eng.pdf;jsessionid=D5EF0D27BE1F21742EF7BDECF713C286?sequence=1.

[112] *Id.* art. 6, para. 1.

[113] Riva Richmond, *Facebook's New Way to Combat Child Pornography*, N.Y. TIMES (May 19, 2011), https://gadgetwise.blogs.nytimes.com/2011/05/19/facebook-to-combat-child-porn-using-microsofts-technology/?partner=rss&emc=rss [https://perma.cc/SE5U-GQWV].

Using the proposed *sua sponte* powers described earlier, the Oversight Board could issue an opinion requiring Facebook to apply computer vision solutions like PhotoDNA to the online illegal wildlife trade, and stating that Facebook must have *proactive* safeguards in place for public health issues.

5. Morals

Morals are perhaps the most ambiguous, and thus troubling, of "legitimate interests." General Comment No. 34 notes that in the previous General Comment No. 22, "the concept of morals derives from many social, philosophical and religious traditions; consequently, limitations . . . for the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition."[114] The HRC goes on to say that "[a]ny such limitations must be understood in the light of universality of human rights and the principle of non-discrimination." The HRC's guidance on morals is so vague as to be meaningless and even dangerous, allowing all kinds of interests to be smuggled through and be considered "legitimate." UNGP 12 provides some clarity, noting that,

> [d]epending on circumstances, **business enterprises may need to consider additional standards** . . . enterprises should respect the human rights of individuals belonging to specific groups or populations that require particular attention, where they may have adverse human rights impacts on them. In this connection, **United Nations instruments have elaborated further** on the rights of indigenous peoples; women; national or ethnic, religious and linguistic minorities; children; persons with disabilities; and migrant workers and their families.[115]

In practice, this would mean that the legitimate interest of morals may be applied with reference to restrictions or obligations on freedom of expression in treaties that are widely adopted (i.e. having many member states). With regards children, for example, the relevant "additional standard" is the Convention on the Rights of the Child (CRC),[116] a treaty ratified by 196 countries. General Comment No. 13 interprets the CRC, with paragraph 21(g) interpreting Article 19 of the CRC ("measures to protect the child from all forms of physical or mental violence"), by stating that child "cyberbullying" is a form of "mental violence" that states need to protect against.[117] Similarly, paragraph 25 of General Comment

[114] General Comment No. 34, *supra* note 76, at para. 32.
[115] *Guiding Principles on Business and Human Rights*, *supra* note 47 (emphasis added).
[116] U.N. Convention on the Rights of the Child, Sept. 2, 1990, 1577 U.N.T.S. 3. https://www2.ohchr.org/english/bodies/crc/docs/CRC.C.GC.13_en.pdf.
[117] UNHRC General Comment No. 13, para. 21(g), CRC/C/GC/13 (Apr. 18, 2011), https://www.refworld.org/docid/4e6da4922.html.

No. 13 details "sexual abuse and exploitation" as something that states should protect against.[118]

Defining "morals" through General Comment No. 22 and GP 12 prevents "morals" from becoming a cover for whatever interests the Oversight Board members might prefer. This is accomplished by limiting the definition of "morals" to whatever has been ratified in multilateral treaties by the global community. It must be noted that definitions of "morals" are still subject to the requirement that Article 19(3) be interpreted consistent with enjoyment of all the other rights in the ICCPR. This means that regressive interpretations of "morals," even those found in other multilateral treaties, would be considered inconsistent and inapplicable.

There have been calls by some commentators to allow national or regional collections of advisors to consult with the Oversight Board on what specific national or regional morals may be applicable to any given case. This is problematic. Relying on the ideologies of a handful of advisors to articulate what "morals" mean gives too much influence to the selection criteria. A better means of articulating morals is to rely on multilateral treaties that have been adopted by a wide array of member states. Any reference to moral interests in the community standards or similar sources of quasi law needs to cite to the applicable multilateral treaty (e.g. CRC) and the relevant principle in the treaty (e.g. protection from child abuse).

C. Adapting Necessity and Proportionality to Social Media Companies

Provided that there is at least one legitimate interest presented by the social media company, the analysis then shifts to a balancing test to assess whether the restriction is necessary and proportionate to protecting that interest. Facebook says it uses a test to balance the voice value of a post against its newsworthiness, public interest, and four additional values: authenticity, safety, privacy, and dignity. In practice, however, as Kettemann and Schulz have documented, Facebook's policy teams do not apply this test in a structured or transparent way.[119] The balancing test I propose, via the necessity and proportionality prong of ICCPR Article 19(3), would bring transparency and structure to Facebook's balancing

[118] *Id.* at para. 25. Note that other sources of IHRL are not only limited to the legitimate interest of morals. In the UN General Assembly Resolution 68/167, for example, "[t]he right to privacy in the digital age," can apply, though that may apply under the rights and reputations of others legitimate interest. *See* G.A. Res. 68/167 (Dec. 18, 2013). The question of which interest applies likely hinges on the distinction between "rights" and "morals." In my view, "rights" are clearly stated as "rights" in sources of IHRL, whereas morals are not. However, there may be enough ambiguity here to comprise an open question for the Oversight Board to decide.

[119] KETTEMANN & SCHULZ, *supra* note 1, at 19.

test and close the gap between what Facebook has promised to do and what it currently does.

The balancing test is by nature a fact-based analysis. The HRC in General Comment No. 34, paragraphs 33-34, relies on HRC decisions; it makes sense that the HRC finds its own decisions persuasive. The Oversight Board's proponents have stated that a proportionality test will be used in the Oversight Board's decision making. The issue here is whether the Oversight Board should give persuasive weight to other tribunals or quasi-judicial bodies such as the HRC or European Court of Human Rights (ECtHR) when making its decisions. As I will detail further below, I believe the answer to this is no.

In applying necessity and proportionality to social media companies, the distinction between private sector social media companies and states becomes profound. Entities like the Facebook Oversight Board have discretion to issue decisions that may have different outcomes than a tribunal adjudicating nation-state issues. The term "least restrictive means" has a different meaning for a social media company than a state, as the latter has the power (i) to issue monetary fines, (ii) to order an actor to do or not do some activity (e.g. injunctive relief), and/or (iii) to deprive someone of their physical freedom (e.g. imprisonment). Social media companies currently have none of these powers–though this is likely to change in the future with the proliferation of e-payments on Facebook's platform.[120] Currently, a company like Facebook only has the power (i) to turn off virality for the content, (ii) to take down the content, (iii) to temporarily ban a user for a period of time, or (iv) to permanently ban a user. As noted by David Kaye, "in each case, it would remain essential for the [social media company] to demonstrate the necessity and proportionality of taking action, *and the harsher the penalty, the greater the need for demonstrating strict necessity*."[121]

D. Necessity and Proportionality: Social Media Companies Are Different

In order to make sure that restrictions on expression are necessary and proportionate, I argue that social media companies should apply an adapted Rabat test. Although Rabat was written to clarify which expressions states should be *required* to prohibit, an adapted Rabat can instead

---

[120] *See* Kevin Webb, *Facebook's New Payment Service Will Let You Send Money Without Fees Across Facebook, Instagram, WhatsApp, and Messenger*, BUSINESS INSIDER (Nov. 12, 2019), https://markets.businessinsider.com/news/stocks/facebook-pay-payments-instagram-whatsapp messenger-send-money-2019-11-1028682585?fbclid=iwar176yjvkwqohp5wmfn43hspedacyx sqlke9vrkj30xryqd2yorhu38w6lu# [https://perma.cc/7KQ7-TS5Q]. For example, Facebook may be able to temporarily or permanently ban accounts from sending or receiving payments, using ad boosts, or even issuing fines for certain infractions. Facebook may also decide to freeze the payment accounts for users who are promoting illegal wildlife trade on Facebook, which could contribute to the spread of zoonotic diseases like Covid-19.

[121] Kaye, *supra* note 22, at para. 20 (emphasis added).

test the *proportionality* of the restrictions that social media companies adopt. As discussed earlier, Rabat is an attempt to clarify ICCPR Article 20(2), which *requires* states to prohibit *by law* "any advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence."[122]

It is important to note that prohibitions can take civil, administrative, or criminal forms.[123] Rabat employs a 6-factor test "to determine the severity necessary to criminalize incitement," and includes the following factors.[124]

- The "**social and political context** prevalent at the time the speech was made and disseminated."
- The **status of the speaker**, "specifically the individual's or organization's standing in the context of the audience to whom the speech is directed."
- **Intent**, meaning that "negligence and recklessness are not sufficient for an offence under article 20 of the Covenant", which provides that mere distribution or circulation does not amount to advocacy or incitement.
- **Content and form of the speech**, in particular "the degree to which the speech was provocative and direct, as well as the form, style, nature of arguments deployed."
- Extent or reach of the speech act, such as the "**magnitude and size of its audience**", including whether it was "a single leaflet or broadcast in the mainstream media or via the Internet, the frequency, the quantity and the extent of the communications, whether the audience had the means to act on the incitement."
- Its **likelihood, including imminence**, meaning that "some degree of risk of harm must be identified", including through the determination (by courts, as suggested in the Plan of Action) of a "reasonable probability that the speech would succeed in inciting actual action against the target group."[125]

While these six tests were designed in order to determine which expression states should be required to prohibit, they can also be relevant to determining the appropriate type of restriction that companies can use when they choose to limit expression on their platforms. David Kaye has already made this connection:

[122] ICCPR, *supra* note 19, art. 20(2).
[123] Rabat, *supra* note 66, at para. 20.
[124] Kaye, *supra* note 22, at para. 14 (citing Rabat, *supra* note 66, at para. 29).
[125] Rabat, *supra* note 66 at para. 29(f).

> **A set of factors is identified in the Rabat Plan of Action that is applicable to the criminalization of incitement under article 20 (2) of the Covenant, but those factors should have weight in the context of company actions against speech as well.** They need not be applied in the same way as they would be applied in a criminal context. However, they offer a valuable framework for examining when the specifically defined content–the posts or the words or images that comprise the post–merits a restriction.[126]

To apply Kaye's suggestion to social media companies, the Rabat test should be part of a floor of content that social media companies should be *required* to prohibit through their own quasi-law (e.g. the Facebook Community Standards), just as Article 20 and Rabat are intended for states. For content regulation that falls outside of Article 20's domain, Article 19(3) should be used, with perhaps a pathway for Rabat and Article 19(3) to assess where on the sliding scale of available punishments the Oversight Board should land. This pathway will be discussed below.

Given that social media companies are different in character from states, regional and domestic tribunal decisions and laws should not have persuasive or precedential weight on decisions made by the Facebook Oversight Board as it relates the criteria of necessity and proportionality. Were this not the case, countries would be able to argue that their domestic laws on freedom of expression should be controlling on the Oversight Board. Observers such as the International Commission of Jurists (ICJ) have noted that domestic legislation on freedom of expression tends to be overbroad and too favourable toward government interpretations of freedom of expression.[127] The ICJ has specifically documented how "Southeast Asian governments have, for decades, crafted and enforced the law to curtail expression and information" with many of these legal frameworks sharing the following characteristics: "vague, overbroad legal provisions; severe and disproportionate penalties; lack of independent oversight mechanisms; and failure to provide effective remedy or accountability."[128]

The ICJ notes that "[c]onceptions of 'national security' and 'public order' have been conflated with the perceived interests of the ruling government or other powerful interest groups to target specific expression. Emerging laws claim extraterritorial application, and in some cases, seek to extend their reach beyond public expression, to private communications. These frameworks either do not advance legitimate aims or do not do so in accordance with applicable principles of legitimacy or necessity

[126] Kaye, *supra* note 22, at para. 49 (emphasis added).
[127] Kaye, *supra* note 22, at para. 4.
[128] INTERNATIONAL COMMISSION OF JURISTS, *Southeast Asia: ICJ Launches Report on Increasing Restrictions on Online Speech* (Dec. 11, 2019), https://www.icj.org/southeast-asia-icj-launches-report-on-increasing-restrictions-on-online-speech [https://perma.cc/WKS8-F9EQ].

and proportionality, and are thus 'in violation of international law.'"[129] For example, Singapore recently passed a "fake news" law that INGO ARTICLE 19 notes gives the "government the power to decide what is true and false in Singapore."[130] Adherence to regional and domestic tribunals and laws, which are projections of state power, are not sound guides for the Facebook Oversight Board. The HRC is right to conclude, in General Comment No. 34, that the scope of freedom of expression "is not to be assessed by reference to a 'margin of appreciation'" as the European Court of Human Rights uses.[131] Instead "a State party, in any given case, must demonstrate in specific fashion the precise nature of the threat to any of the enumerated grounds listed in paragraph [Article 19](3) that has caused it to restrict freedom of expression."[132]

The discussion above suggests that it is best to allow the Oversight Board to determine the "least restrictive means" that are "directly related" to the specific legitimate interest at hand. The Oversight Board is likely to issue decisions through trial and error on virality switch offs, takedowns, temporary and permanent bans, and other means that social media companies have in contrast to the sanctioning powers nation states have. This sliding scale will be referenced against a growing "case law" of borderline fact patterns that will adumbrate what necessity and proportionality mean as applied to social media companies.[133]

Take, as a hypothetical, an initial decision involving a Facebook post by a head of state inciting violence against Muslims. The Oversight Board runs the facts of this case through the Rabat 6-factor test and finds the following:

1) **Social and political context**: for this country there is a recent history of actual violence against Muslims. On a severity scale from 1 (least severe) to 5 (most severe), the Board gives a score of 4.
2) **Status of the speaker**: the speaker is a head of state, so the Board gives a score of 5.
3) **Intent**: the head of state claims that he was joking when he made the comment, so the Board gives a score of 2.

---

[129] *Id.*

[130] ARTICLE 19, *Singapore: New Law on 'Online Falsehoods' a Grave Threat to Freedom of Expression* (May 9, 2019), https://www.article19.org/resources/singapore-new-law-on-online-falsehoods-a-grave-threat-to-freedom-of-expression.

[131] General Comment No. 34, *supra* note 76, at para. 36.

[132] General Comment No. 34, *supra* note 76, at para. 36.

[133] *See Oversight Board Bylaws* § 1.2.1, FACEBOOK (Jan. 2020), https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf ("The case selection committee will set criteria (e.g. importance and precedential impact) for the cases that the board will prioritize and select for review, which may change over time"); *Oversight Board Charter,* art. 2, § 2, FACEBOOK ("For each decision, any prior board decisions will have precedential value and should be viewed as highly persuasive when the facts, applicable policies, or other factors are substantially similar.").

4) **Content and form of the speech**: the speaker said "all Muslims want to do is take over our country. Every good Christian should get a gun and kill them all." The Board gives a score of 5.
5) **Magnitude and size of its audience**: the speaker has a large following on Facebook, with over 20 million fans. The post has 50,000 likes, 10,000 shares, and 2 million views. The Board gives a score of 4.
6) **Likelihood (including imminence)**: the speaker did not speak precisely as to the time, date, and venue, but did mention the use of guns, in a country with previous killings of Muslims. Christians have said that they were inspired by the president. The Board gives a score of 4.

With an overall score of 24 out of 30,[134] the Board finds the speech severe enough to ban the post despite the head of state's comments being "newsworthy." A subsequent Board decision may involve a private individual Facebook post stating that the individual "hates Muslims," which is shared with the individual's friends. Were the Board, in applying the same Rabat 6-factor test, to permanently ban this user, the Board would likely be in violation of its precedent given that the overall scores for this woman's speech are likely to be lower than 24 out of 30.

This numerical scoring system ensures that the severity of the speech is taken into account in determining the appropriate sanction, thus satisfying the necessity and proportionality test of Article 19(3). All factors need not be present in any given fact pattern, as the intensity/severity of each individual factor can vary greatly. For example, perhaps "intent and imminence" factors are not found, in which case a milder punishment may be recommended–though a *punishment can still be imposed.* This is unlike the Rabat test as it was developed for Article 20(2), which would only necessitate a punishment if all six factors were satisfied.

Entities like the Oversight Board may consider this adapted Rabat test or come up with a similar (or different) test. Such tests can also be used to assess whether the algorithms Facebook deploys to limit or ban content are using the "least restrictive means" when rendering sanctions. If, for example, the Oversight Board finds that the algorithms are trained on a mislabelled data set to remove false positives by applying the Rabat test (e.g. the algorithm overweights the calculation of speaker status and intent), it could issue a decision requiring Facebook to more narrowly tailor its algorithm and training data labels to reduce such takedowns.

[134] Note that I am not suggesting the Board use a numerical scoring system in practice–the numbers are used to clarify the example for the reader.

E. Applying this Framework to Facebook Oversight Board Decisions

Imagine an edge case is brought before the Facebook Oversight Board. This case concerns whether Facebook's takedown of a coronavirus post, which features a video clip of a country's president claiming that the novel coronavirus is a Chinese government bioweapon run amok, is an appropriate content moderation decision.

First, an Article 20 analysis is conducted. Under my formulation of Article 20, which states that intense and irrational public speech that intends to incite discrimination, hostility, or violence based on nationality, race, or religion should be blocked, there does not appear to be clear evidence of intent in this case. This means that Facebook is not *required* to ban the content.

Facebook, however, *may* ban the content, and so the proposed framework shifts to an Article 19(3) analysis which examines (i) legality, (ii) legitimacy, and (iii) necessity and proportionality. Under the legality prong, Facebook argues that its decision to take down the content is in line with its Community Standards on "False News."[135] Facebook's Newsroom post on coronavirus states that when "[o]ur global network of third-party fact-checkers . . . rate information [related to the coronavirus] as false, we limit its spread on Facebook and Instagram and show people accurate information from these partners. We also send notifications to people who already shared or are trying to share this content to alert them that it's been fact-checked. We will also start to remove content with false claims or conspiracy theories that have been flagged by leading global health organizations and local health authorities that could cause harm to people who believe them."[136] The Oversight Board finds that Facebook's False News Community Standards are not "precise" as required by the legality prong of Article 19(3), since they are silent on public health matters. They find the Newsroom post to not be precise or transparent in terms of how its "global network of third-party fact-checkers" decide such content is false. Moreover, they take issue with Facebook's failure to determine who its "leading global health organizations and local health authorities" are.[137] However, they do find the standards to be "precise, public and transparent" in terms of limiting virality.[138]

The Oversight Board directs Facebook to articulate its fact-checking process and identify, by name, its "leading global health organizations and local health authorities." The Board requires Facebook to disclose the computer science labels that its content reviewers use to label posts

---

[135] *False News*, FACEBOOK, https://www.facebook.com/communitystandards/false_news (last visited July 21, 2020).
[136] Jin, *supra* note 108.
[137] Jin, *supra* note 108.
[138] Kaye, *supra* note 22, at para. 6(a).

for training data purposes, and how it designs its algorithms to automate this human cognitive work. The Oversight Board notes that Facebook had no "public comment" process for banning coronavirus false news, though it acknowledges that there are "expedient circumstances"[139] in light of the pandemic that may warrant Facebook taking action before opening up a public comment period. The Oversight Board notes that Facebook has "unbounded discretion" to decide this; however, it requires that it get approval from the World Health Organization (WHO) in the future before issuing new rules related to public health, unless the WHO unduly delays its approval and Facebook is compelled to act quickly due to expedient circumstances, with such approval eventually to be obtained in due course. The Oversight Board notes that the Board's own existence, plus the required input from the WHO, would be good "procedural safeguards."

Under the legitimacy prong, Facebook claims it has both "public health" and "public order" interests in preventing harm. Facebook also claims that the "public health" interest requires them to give notice to the WHO of COVID-19 misinformation and disinformation on the platform. The Oversight Board agrees with Facebook's public health claims; however, since the President only said China could "not be trusted" rather than inciting some disruption or violent action against China, the "public order" interest is not found.

Finally, under the necessity and proportionality prong, the Oversight Board finds Facebook's decision to turn off virality for false news posts and instead show correct information to be the "least restrictive means" to protect public health interests. After Facebook provides the algorithmic labels to the Oversight Board,[140] the Oversight Board, with assistance from its technical staff,[141] concludes that the algorithmic weights for punishment recommendations are not aligned with the degree of severity for each of the Rabat test's six factors. The Oversight Board finds that Facebook's decision to ban posts that only signal distrust (instead of, say, inciting physical harm) to be not the "least restrictive means" available. The Oversight Board therefore instructs Facebook to abandon the takedowns, and instead to only turn off virality for these posts and to

[139] This would be an example of the Oversight Board having the flexibility of making new "common law" for the peculiar nature of ever-evolving social media companies. The proposed IHRL framework in this essay is very permissive of the Oversight Board to come up with new rules; the essay merely proposes a reasoning framework rooted in IHRL to serve as a foundation and independent check for Oversight Board decisions that has widespread legitimacy among the global community due to overwhelming adoption of multilateral treaties like the ICCPR and CRC by most countries of the world.

[140] Facebook here may require that the labels and algorithms provided to the Oversight Board being kept under seal, for risk of exposing its processes to coordinated propagandists who could learn how to evade detection. This essay does not have an opinion as to whether such seals should be given and leaves this question open for other commentators to decide.

[141] Facebook has told the author that the support staff for the Oversight Board will include people with technical backgrounds in computer science.

point to corrections written by Poynter-approved fact-checkers.[142] Facebook must also tweak its algorithms and labels to ensure that this change is implemented in its automated detection of coronavirus misinformation moving forward. The Oversight Board's technical staff, using encrypted enclaves as part of the Confidential Computing Consortium,[143] audits Facebook's algorithms[144] to ensure that they have been tweaked properly in accordance with this decision.

The example above is obviously hypothetical. The Oversight Board may not rule in this fashion. That being said, my intent is to show how the Oversight Board can employ a structured decision-making procedure that not only accords with IHRL, but also gives the Oversight Board the flexibility it needs to come up with new "common law" that addresses the novel context of social media companies.

V. Conclusion

The Facebook Oversight Board has the potential to become an immense, world-changing success. It could constitute a new quasi-judicial body that draws on a respected source of law, the IHRL, to which most states have already consented, and advance freedoms of expression and the protection of certain classes of people such as racial and ethnic minorities, women, and children. Crucially, the Oversight Board and other Social Media Councils would *benefit* from adopting IHRL because it would give these entities a basis for pushing back against illegitimate orders or requests from nation states using the IHRL framework.

The Oversight Board would give international law teeth and serve as an antidote to the growing geopolitics enveloping tech, especially between China and the United States. If successful, the Oversight Board could turn into a more general Social Media Council, with all social media companies coming under its ambit, regardless of national origin. At the same time, the Oversight Board could also become what its detractors fear: a kangaroo court.

Forthright adoption and adaptation of international human rights law is the key to giving the Oversight Board a sense of legitimacy in its fledgling days, especially as it struggles to prove that it is more than just a cover for Facebook to continue business as usual. I hope that this paper can act as a guidepost for how IHRL can be usefully applied in the Oversight Board's decisions. By providing a functional framework that grants the Oversight Board discretion to narrowly tailor decisions, Article 19 of

---

[142] POYNTER: THE INTERNATIONAL FACTCHECKING NETWORK (IFCN), https://www.poynter.org/ifcn(last visited Jul. 21, 2020). The Poynter IFCN is considered by the UN and other agencies to be the leading fact-checking institute in the world.

[143] CONFIDENTIAL COMPUTING CONSORTIUM, MEMBERS, https://confidentialcomputing.io/members (last visited Jul. 21, 2020).

[144] OPEN ENCLAVE SDK, https://openenclave.io/sdk (last visited Jul. 21, 2020).

the ICCPR can allow the Oversight Board to come up with its own "common law." This framework is not the final word, but rather the start of a robust debate and discussion about how the Oversight Board can make content decisions and policy recommendations while respecting human rights.