2020

But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies

Susan Benesch

But Facebook's Not a Country: How to Interpret Human Rights Law for Social Media Companies

Susan Benesch†

Private social media companies regulate much more speech than any government does, and their platforms are being used to bring about serious harm. Yet companies govern largely on their own, and in secret.

To correct this, advocates have proposed that companies follow international human-rights law. That law–by far the world's best-known rules for governing speech–could improve regulation itself, and would also allow for better transparency and oversight on behalf of billions of people who use social media.

This paper argues that for this to work, the law must first be interpreted to clarify how (and whether) each of its provisions are suited to this new purpose. For example, the law provides that speech may be restricted to protect national security, as one of only five permissible bases for limiting speech. Governments, for which international law was written, may regulate on that basis, but not private companies which have no national security to protect.

To fill some of the gap, the paper explains and interprets the most relevant provisions of international human-rights law–Articles 19 and 20 of the International Covenant on Civil and Political Rights, which pertain to freedom of expression–for use by social media companies, in novel detail.

I. Introduction

Facebook, Inc. is quietly running the largest system of censorship the world has ever known, governing more human communication than any government does, or ever has[1]: billions of posts every day.[2] Other social

† Executive Director of the Dangerous Speech Project and Faculty Associate, Berkman Klein Center for Internet & Society, Harvard University. I am very grateful to Chinmayi Arun, Dan Bateyko, Matthew Bugher, Cathy Buerger, Evelyn Douek, Tonei Glavinic, David Kaye, Michael Lwin, Yuval Shany, and Miranda Sissons for invaluable discussions and critique, and to Dan Bateyko, Tonei Glavinic, and Emma Hunter for research and editing.

[1] The only country whose censorship system comes close to that of Facebook's in number of users or volume of content regulated is China, which has fewer than one billion people online. *See, e.g.*, Jon Russell, *China Reaches 800 Million Internet Users,* TECHCRUNCH (Oct. 21, 2018), https://techcrunch.com/2018/08/21/china-reaches-800-million-internet-users/ [https://perma.cc/WZ98-PSZN]. Facebook has more than 2.3 billion regular monthly users. *See* FACEBOOK, https://newsroom.fb.com/company-info/ [https://perma.cc/5FHJ-QJ8W].

media companies[3] also regulate speech at an astounding scale, on platforms[4] like YouTube.[5]

It's not only scale that makes this form of regulation new. In the platforms' early days nearly all posts were private: users shared personal news and thoughts, often with people they knew offline. Significant public discourse now takes place on social media, among strangers. In many countries, online platforms are indispensable for running election campaigns, announcing presidential decisions, providing government services[6]–or convincing citizens to condone mass murder and ethnic cleansing.[7]

Though the vast majority of online content is harmless or helpful, platforms are also used to spread many types of content that do significant or grave harm to people, and that damage the public good. The link between social media company decisions and real-world suffering is not as obvious, and the solution not as easy, as in cases of products that kill people directly, like the Ford Pinto with its exploding gas tank,[8] or HIV-contaminated clotting factor knowingly sold to hemophiliacs.[9]

However, the harm is real. When companies seek profit by maximizing user engagement and retention, users can be, and have been, exposed to damaging content that leads to offline harm.[10]

---

[2] "Users post billions of pieces of content to our platform every day and send hundreds of billions of messages through our chat services." Email from Peter Stern, Facebook Director of Content Policy Stakeholder Engagement, to Susan Benesch (Aug. 19, 2020, 5:46PM EST) (on file with author).

[3] In this paper, "social media companies" refers to firms that host and disseminate user-generated content online. Facebook, Google, and Twitter are the best known and most discussed (at least in the United States). There are many others, large and small, including Reddit, Automattic, Bytedance, and companies that build and maintain chat apps or online games.

[4] In this paper "platform" is any product, app, or website on which a social media company hosts and disseminates user-generated content. For example, Facebook, Inc., a social media company, owns and operates the platforms Facebook, Instagram, and WhatsApp.

[5] YouTube has nearly two billion users, and sees more than 400 hours of video posted every minute. YouTube Help, *Monetization Systems or "The Algorithm" Explained*, GOOGLE INC., https://support.google.com/youtube/answer/9269689 [https://perma.cc/NH4E-533S].

[6] Arthur Mickoleit, *Social Media Use by Governments: A Policy Primer to Discuss Trends, Identify Opportunities, and Guide Decision Makers*, OECD 2 (2014), https://read.oecd-ilibrary.org/governance/social-media-use-by-governments_5jxrcmghmk0s-en#page1 [https://perma.cc/9WHP-FFA9].

[7] *See, e.g.*, Alexandra Stevenson, *Facebook Admits It Was Used to Incite Violence in Myanmar*, N.Y. TIMES (Nov. 6, 2018), https://www.nytimes.com/2018/11/06/technology/myanmar-facebook.html [https://perma.cc/GC8W-SGYL]; Davey Alba, *How Duterte Used Facebook to Fuel the Philippine Drug War*, BUZZFEED NEWS (Sept. 4, 2018), https://www.buzzfeednews.com/article/daveyalba/facebook-philippines-dutertes-drug-war [https://perma.cc/YW6C-DKJ2].

[8] Robert Sherefkin, *Lee Iacocca's Pinto: A Fiery Failure*, AUTOMOTIVE NEWS (June 16, 2003), https://www.autonews.com/article/20030616/SUB/306160770/lee-iacocca-s-pinto-a-fiery-failure [https://perma.cc/86HM-97M3].

[9] Leemon McHenry & Mellad Khoshnood, *Blood Money: Bayer's Inventory of HIV-Contaminated Blood Products and Third World Hemophiliacs*, 21 ACCOUNTABILITY IN RESEARCH 389-400 (2014), https://www.ncbi.nlm.nih.gov/pubmed/24785997 [https://perma.cc/UYC2-F84R].

[10] Stevenson, *supra* note 7; Alba, *supra* note 7.

Yet the companies decide–many millions of times every day[11]–which content to regulate and how, such as by removing it,[12] posting warning notices,[13] fact-checking,[14] or making it visible to fewer people by means of algorithms.[15] They choose which accounts to suspend or block entirely, and which to leave up so that their owners can continue exercising freedom of expression, even when that infringes on the rights of others.

For all this regulation the companies make their own rules, frequently changing them, on their own initiative or in response to public pressure, and keeping the detailed rules[16] secret from the people whose speech is regulated under their terms. Sometimes company staff admit later that they erred,[17] either by leaving content up or by taking it down. In the last several years, under pressure to regulate more effectively, platforms have hired thousands more moderators, developed software to detect harmful content, and constantly adjusted their rules or made exceptions to them.[18]

Each platform has its own regulations, such as Facebook's Community Standards,[19] which many users do not even know how to find–and most do not read.[20]

---

[11] Facebook alone restricts access to over one million pieces of content every day, not including spam. *See Community Standards Enforcement Report*, FACEBOOK (May 2020), https://transparency.facebook.com/community-standards-enforcement [https://perma.cc/28QT-3H8H].

[12] *See, e.g.*, *Community Standards*, FACEBOOK, https://facebook.com/communitystandards [https://perma.cc/7VAX-BP3H].

[13] *See, e.g.*, Help Center, *Sensitive Media Policy*, TWITTER, (Nov. 2019), https://help.twitter.com/en/rules-and-policies/media-policy [https://perma.cc/XY24-JB9F].

[14] *See, e.g.*, Tessa Lyons, *Hard Questions: What's Facebook's Strategy for Stopping False News?*, FACEBOOK (May 23, 2018), https://about.fb.com/news/2018/05/hard-questions-false-news/ [https://perma.cc/G52V-S6FT].

[15] *See, e.g.*, Arun Babu, Annie Liu & Jordan Zhang, *New Updates to Reduce Clickbait Headlines*, FACEBOOK (May 17, 2017), https://about.fb.com/news/2017/05/news-feed-fyi-new-updates-to-reduce-clickbait-headlines/ [https://perma.cc/U9VG-9T42].

[16] *See* Section II.B *infra* for an explanation of the difference between public-facing guidelines like Facebook's Community Standards and the detailed, internal rules that companies use for regulating.

[17] *See, e.g.*, Shona Ghosh, *Google Admitted Its AI Is Making Errors When Deleting Terrorist Content Off YouTube*, BUS. INSIDER (Oct. 18, 2017, 6:43AM EST), https://www.businessinsider.com/google-errors-delete-terrorist-videos-youtube-researchers-journalists-2017-10 [https://perma.cc/KZ37-7XFK]; Joachim Dagenborg, *Facebook Says Will Learn from Mistake over Vietnam Photo*, REUTERS (Sept. 12, 2016, 12:10 PM EST), https://www.reuters.com/article/us-norway-facebook-idUSKCN11I1VU [https://perma.cc/QFX8-YG79]; Ariana Tobin, Madeleine Varner & Julia Angwin, *Facebook's Uneven Enforcement of Hate Speech Rules Allows Vile Posts to Stay Up*, PROPUBLICA (Dec. 28. 2017, 5:53 PM EST), https://www.propublica.org/article/facebook-enforcement-hate-speech-rules-mistakes [https://perma.cc/8PTN-RLPR].

[18] *See, e.g.*, Susan Wojcicki, *Expanding Our Work Against Abuse of Our Platform,* YOUTUBE OFFICIAL BLOG (Dec. 5, 2017), https://blog.youtube/news-and-events/expanding-our-work-against-abuse-of-our [https://perma.cc/EB4B-9MCG]; Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1648-58 (2018).

[19] *Community Standards*, *supra* note 12.

[20] Jonathan A. Obar & Anne Oeldorf-Hirsch, *The Biggest Lie on the Internet: Ignoring the Privacy Policies and Terms of Service Policies of Social Networking Services*, 23 INFO., COMM. &

This, of course, makes users less likely to follow the rules.[21] And those who do read them are often confused: even some company staff find them difficult to understand or apply clearly.[22]

Thus an external source of standards is needed. The companies' own rules and processes are not only confusing and obscure to users, but also unevenly enforced (more in some languages and countries than others), and subject to interference from governments. In the words of David Kaye, then United Nations Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression, they have "created unstable, unpredictable, and unsafe environments for users and intensified government scrutiny."[23]

Adhering to a single body of standards would instead give companies a source of "forceful normative responses against undue State restrictions"[24] and would provide a basis for users to hold companies accountable to follow key norms and principles. Also as ARTICLE 19, the NGO that champions freedom of expression worldwide, argues, applying human-rights law to content moderation would oblige companies to disclose more information about rules and enforcement, and might lead them to provide more effective remedies to users who feel companies have violated their rights.[25]

Social media companies need a clear, authoritative, widely endorsed source of parameters for rules–and of guidance on how to regulate, including disclosing and explaining their rules to the public. No source of rules for speech regulation is as widely known or formally adopted as international human-rights law. That law's most relevant instrument, the International Covenant on Civil and Political Rights (ICCPR), has been ratified by nearly ninety percent of the countries in the world.[26] It should be noted that this does not mean they comply with the treaty's terms. The United States, for example, ratified with a wall of fine print–five reserva-

---

SOC'Y 128-47 (2020); *Id.*, 44th Research Conference on Communication, Information & Internet Policy (2016) https://ssrn.com/abstract=2757465 [https://perma.cc/K8ZU-N3L3].

[21] Margaret E. Tankard & Elizabeth Levy Paluck, *Norm Perception as a Vehicle for Social Change*, 10 SOC. ISSUES & POL'Y REV. 181 (2016).

[22] STEVEN LEVY, FACEBOOK: THE INSIDE STORY, 447-48, (2020).

[23] David Kaye (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, 14, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018).

[24] *Id.*

[25] *Side-Stepping Rights: Regulating Speech by Contract*, ARTICLE 19, at 36 (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf [https://perma.cc/NQ9P-PWDH].

[26] United Nations Treaty Collection, *Status of Treaties*, UNITED NATIONS (Aug. 6, 2020 04:27 EDT), https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-4&chapter=4&clang=_en [https://perma.cc/W5T4-ZNYH].

tions, five understandings, and four declarations[27]–to curb its own compliance.

Companies are enjoined to respect human-rights law (without reservations), under the UN Guiding Principles on Business and Human Rights which establish a "global standard of expected conduct."[28] Companies like Facebook should not only respect the law but comply with it, due to their vast regulation of speech, especially public discourse, Kaye argued in a report to the UN Human Rights Council: "While the Guiding Principles are non-binding, the companies' overwhelming role in public life globally argues strongly for their adoption and implementation."[29] ARTICLE 19 has also called for this extensively.[30]

Companies need not–and should not–all use the same rules, just as countries are not required to pass the same national laws. Most social media platforms are used by a wide variety of people in many countries, for diverse purposes that call for different rules. Fortunately, with only two exceptions, the ICCPR does not dictate which speech should be restricted.[31] Instead it sets out a process: a regulatory framework to make sure the rules are transparent, designed to protect the relevant populations, and minimally restrictive of freedom of expression.

However, the ICCPR and other relevant human-rights law cannot be used for commercial content moderation right off the shelf; it must be interpreted for this novel purpose in at least two ways. First, the law was written and ratified for use by countries, not private companies. The ICCPR allows speech restrictions on only five grounds–national security, public order, public health, morals, and the rights and reputations of others. No other grounds are permitted[32]– profit-making for example. As Evelyn Mary Aswad has argued, companies prioritize profit, and perhaps cannot be expected to put the public interest above it.[33]

It would be naïve to expect companies to make decisions *only* in the public interest, but in other industries, businesses have been obliged to make some decisions to protect consumers, their workers, the environ-

---

[27] 138 CONG. REC. S4781-01 (daily ed. Apr. 2, 1992) (U.S. Reservations, Declarations, and Understandings, International Covenant on Civil and Political Rights) http://hrlibrary.umn.edu/usdocs/civilres.html [https://perma.cc/5839-6TX8].

[28] UN, *The Corporate Responsibility to Respect Human Rights: An Interpretive Guide*, U.N. Doc. HR/PUB/12/02 at 13 (2012) https://www.ohchr.org/Documents/publications/hr.puB.12.2_en.pdf [https://perma.cc/4V8U-QM3V]

[29] Kaye, *supra* note 23, at 5.

[30] ARTICLE 19, *supra* note 25 at 9-10, 39.

[31] Article 20 of the ICCPR requires states to prohibit propaganda for war and advocacy of national, racial, or religious hatred that constitutes incitement to hostility, discrimination or violence, as discussed in Section III *infra*. G.A. Res. 2200A (XXI), International Covenant on Civil and Political Rights (Dec. 16, 1966), https://www.ohchr.org/en/professionalinterest/pages/ccpr.aspx [https://perma.cc/2RQ6-6EWN] [hereinafter ICCPR].

[32] *Id.* at art. 19.

[33] Evelyn Mary Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 52-53 (2018), https://scholarship.law.duke.edu/dltr/vol17/iss1/2/ [https://perma.cc/ES6W-HBW8].

ment, or the public, by means of oversight mechanisms and law. Social media companies should be no exception. Some provisions of human rights law on speech cannot apply to private enterprise, however, as described below.

The second area in which international human rights law must be interpreted for use by social media companies is where it *requires* speech restrictions, since that language is broad and confusing. The ICCPR's prohibitions of propaganda for war and incitement to discrimination, hostility, or violence must be explained for application to content moderation, which requires clear, specific rules for consistent decisions on a huge variety of posts in over a hundred languages[34] and many more cultural contexts. Moreover, another international human rights treaty requires that other categories of speech be prohibited, including "all dissemination of ideas based on racial superiority or hatred."[35] This seems considerably broader than the ICCPR's speech provisions, and the two treaties should be explicitly reconciled.

If and when it is adequately interpreted, international human-rights law might serve as a useful guide for companies to regulate speech – and for outsiders to hold the companies accountable to an external standard. This should improve the companies' rules, since the law requires that speech restrictions be necessary, legitimate, and provided by law.[36] Users would be better able to understand the rules, and to hold companies accountable for errors.

Companies should indeed conform to international human-rights law, not merely respect it. To that end, below I offer many specific ideas for interpreting that law for use by private companies and by external bodies that are beginning to play a role in the companies' regulation of speech, such as Facebook's new Oversight Board,[37] which could be a mechanism for improving Facebook's speech regulation. And as Evelyn Douek has pointed out, importantly, companies have begun to regulate together in what she calls "content cartels,"[38] like the Global Internet Fo-

---

[34] Maggie Fick & Paresh Dave, *Facebook's Flood of Languages Leave It Struggling to Monitor Content*, REUTERS (Apr. 23, 2019, 1:50 AM), https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leave-it-struggling-to-monitor-content-idUSKCN1RZ0DW [https://perma.cc/AD6A-DMJ5].

[35] United Nations International Convention on the Elimination of All Forms of Racial Discrimination, Dec. 21, 1965, 660 U.N.T.S. https://ohchr.org/en/professionalinterest/pages/cerd.aspx, [https://perma.cc/2EEV-NYWH] [hereinafter ICERD].

[36] General Comment No. 34, at ¶ 24, 33, U.N. Human Rights Comm'n, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011) https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf [https://perma.cc/2894-AKLM] [hereinafter GC 34].

[37] The Oversight Board was proposed in 2018 and is expected to begin work this year. Michael Lwin discusses the Board in valuable detail in his own article. *See Oversight Board*, FACEBOOK (Sept. 2019), https://www.oversightboard.com/ [https://perma.cc/2Z3Q-K2EH].

[38] Evelyn Douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INST. AT COLUM. U. (Feb. 11, 2020), https://knightcolumbia.org/content/the-rise-of-content-cartels [https://perma.cc/882J-J3N4].

rum to Counter Terrorism (GIFCT), a consortium of tech firms that maintains a database of content to ban.[39]

Interpreting international human-rights law for use by social media companies is a major endeavor, which should consider multiple sources of law, and many rights. In this brief essay, I focus on freedom of expression and on one law, the ICCPR.

II. Social Media Companies Should Comply with International Human-Rights Law

A. Social Media Companies Are Governing: Performing Some Limited State Functions

International human-rights law was meant to bind states, but the idea of holding corporations accountable under that law is far from new: it emerged when other companies gained unusual forms or degrees of power over the exercise of human rights. In the late twentieth century, transnational corporations notoriously violated rights,[40] so the UN commissioned a proposal to require corporations to comply with human-rights obligations.[41] It was never adopted. Instead the UN Secretary General appointed Professor John Ruggie who, after six years of discussions, produced the UN Guiding Principles on Business Human Rights [hereinafter UNGPs], which ask companies to "respect" human rights.[42] The UNGPs were unanimously endorsed by the UN Human Rights Council in 2011 and have become the principal international instrument regarding companies and human rights.[43]

---

[39] GIFCT was founded in 2017 as a partnership of large social media companies, and incorporated as an NGO in 2020 with companies on its governing board and an advisory committee of government and civil society representatives. *See* Chloe Hadavas, *The Future of Free Speech Online May Depend on This Database*, SLATE (Aug. 13, 2020, 9:00 AM EST), https://slate.com/technology/2020/08/gifct-content-moderation-free-speech-online.html [https://perma.cc/SU39-NG5V].

[40] For example, Nike relied on factories that abused workers in many countries and Shell was associated with a military crackdown that left thousands dead in Nigeria. John Gerard Ruggie, *The Social Construction of the UN Guiding Principles on Business and Human Rights* 7-8 (Corp. Responsibility Initiative, Working Paper No. 67, 2017) [https://perma.cc/CK84-NEXJ].

[41] Pini Pavel Miretski & Sascha-Dominik Bachmann, *The UN "Norms on the Responsibility of Transnational Corporations and Other Business Enterprises with Regard to Human Rights": A Requiem*, 17 DEAKIN L. REV. 5, 7 (2012) [https://perma.cc/QET6-66LU].

[42] U.N. Office of the High Commissioner of Human Rights, Guiding Principles on Business and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework, U.N. Doc. HR/PUB/11/04 at 17-18 (2011), www.ohchr.org/Documents/Publications/GuidingPrinciplesBusinessHR_EN.pdf [https://perma.cc/8VWD-J74K] [hereinafter UNGPs].

[43] Mariette Van Huijstee, Victor Ricco & Laura Ceresna-Chaturvedi, *How to Use the UN Guiding Principles on Business and Human Rights in Company Research and Advocacy*, The Centre for Research on Multinational Corporations (Nov. 2012), at 11. https://corporatejustice.org/how-to-use-the-un-guiding-principles-on-business-and-human-rights-in-company-research-and-advocacy.pdf [https://perma.cc/DG23-NP3S]

The UNGPs focus more on process than on outcomes, leaving room for many levels of commitment. They indicate that companies ought to develop human-rights policies, conduct due diligence,[44] "avoid infringing on the human rights of others," and should "address adverse human rights impacts with which they are involved."[45]

Social media companies have human-rights impacts of great magnitude on at least half of the world's population.[46] They can limit not only what people say (or write) but what they see and read. When exercised in public spaces, such powers are reserved for governments. The companies wield them in virtual spaces with many public features: what Yochai Benkler called networked public spheres.[47]

This extraordinary, transnational power and influence sets the companies apart from any other private enterprise. When platforms are used for exchanging information that is vital for civic life, the owners and staff of the platforms influence the political, cultural, and economic development of entire societies. In many countries, social media platforms have become essential spaces for public discourse, with major implications for political life, collective well-being, and public order.[48]

U.S. President Donald Trump is far from the only political figure to communicate and campaign on social media. In the Philippines, for example, President Rodrigo Duterte's election in 2016 was significantly credited to campaigning on Facebook, by far the dominant platform in his country.[49]

The very functions of routine governance are also carried out, increasingly, on social media platforms. Not only political figures but the very institutions of government now communicate and provide services via social media, in many countries. "Presence and activity on social media is no longer a question of choice for most governments," wrote the Organisation for Economic Co-operation and Development (OECD) in a working paper. In most of the 34 OECD countries, executive branches maintain both Facebook and Twitter accounts, as do many agencies and ministries, the paper notes.[50]

In sum, social media companies have acquired more regulatory power of a kind traditionally exercised by states, than any other private enterprise has since human rights law was created. Though the UNGPs are cautious and aspirational, the companies should develop robust, rigorous,

---

[44] UNGPs, *supra* note 43.
[45] *Id.* at 13 (Principle 11).
[46] Simon Kemp, *Digital 2020: 3.8 Billion People Use Social Media*, WE ARE SOCIAL (Jan. 30, 2020), https://wearesocial.com/blog/2020/01/digital-2020-3-8-billion-people-use-social-media [https://perma.cc/42YE-HBLQ].
[47] YOCHAI BENKLER, THE WEALTH OF NETWORKS 261 (2006).
[48] *See, e.g.*, Stevenson, *supra* note 7.
[49] Alba, *supra* note 7.
[50] Mickoleit, *supra* note 6, at 1.

transparent means to respect and comply with international human-rights norms.

Other observers have been noting for years that internet companies govern, without meaningful accountability to the people under their control.

In a landmark 2012 book, Rebecca MacKinnon called the companies "sovereigns operating without the consent of the networked."[51] In some cases, the leaders of companies have themselves noted that they are wielding sovereign powers. Kate Klonick began her 2018 article aptly titled *The New Governors* with a remark from Mark Zuckerberg: "In a lot of ways Facebook is more like a government than a traditional company. We have this large community of people, and more than other technology companies we're really setting policies."[52] Klonick called Facebook and other companies "private, self-regulating entities" that have developed such detailed rules and systems for moderating online content that they constitute systems of governance.[53]

Tarleton Gillespie noted that defining the boundaries of freedom of expression online is "enormous cultural power held by a few deeply invested stakeholders, and it is being done behind closed doors, making it difficult for anyone else to inspect or challenge."[54] Regulating in the dark does significant harm: "[T]he biggest threat this private system of governance poses to democratic culture is the loss of a fair opportunity to participate, which is compounded by the system's lack of direct accountability to its users."[55]

Fortunately, human rights law requires that rules be transparent, so that the people who live under them can understand and challenge them. Applying that law would also confer other benefits, such as giving companies a stronger basis to resist improper pressure from states to suppress speech.

Finally, since social media platforms serve very different purposes, the companies will (and may) implement the law differently. None is as large or powerful as Facebook, and some, such as dating apps or gaming platforms, do not host significant public discourse.

Under the UNGPs all private enterprises of any size or "operational context"[56] have a responsibility to respect international human-rights law, but "[n]evertheless, the scale and complexity of the means through which

---

[51] REBECCA MACKINNON, CONSENT OF THE NETWORKED: THE WORLDWIDE STRUGGLE FOR INTERNET FREEDOM, at xxiii-iv.

[52] Klonick, *supra* note 18, at 1599 (citing DAVID KIRKPATRICK, THE FACEBOOK EFFECT: THE INSIDE STORY OF THE COMPANY THAT IS CONNECTING THE WORLD, 254 (2010)).

[53] *Id.* at 1603.

[54] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA 197 (2018).

[55] Klonick, *supra* note 18, at 1603.

[56] *Corporate Responsibility*, *supra* note 28, at 20.

enterprises meet that responsibility may vary according to these factors and with the severity of the enterprise's adverse human-rights impacts."[57] Accordingly, a platform's size, target audience, or intended use should not determine whether its owner should use ICCPR standards as the basis for content moderation.

The scope and severity of human-rights impacts should be the key factor: when significant harm arises from content spreading on a company's platform or platforms, its obligation to adhere to human-rights law grows accordingly. Though some apps are free from propaganda for war, for example, most have some content that violates human rights. Users banned from one site flee to smaller, niche services like LinkedIn,[58] TikTok,[59] and even Pornhub[60] to continue spreading harmful content. Also, platforms grow and morph quickly. When Mark Zuckerberg and others created Facebook in their dorm rooms,[61] none imagined what it would become. Or Amazon's Twitch livestreaming service was created for gamers, but its largest category is now non-gaming content.[62]

B. Private Regulation is Flawed in Ways that International Human-Rights Law Can Improve

A single substantive basis for content moderation would help to solve several of the fundamental problems of tech companies' regulatory systems for content. In brief summary, these are: 1) that the rules are obscure, confusing, or totally unknown to those who live under them; 2) that companies too often make wrong decisions that cause a variety of forms of serious harm to people including but not limited to their users; 3) that the companies' leaders and staff are imposing their values on more than half of the world's population; and 4) that governments can carry out shadow censorship by putting quiet but vigorous pressure on companies.

---

[57] *Id.* at 18.

[58] Craig Silverman, *The Partisan Meme Wars Have Come for LinkedIn,* BUZZFEED NEWS (Nov. 5, 2018, 4:42 PM EST), https://www.buzzfeednews.com/article/craigsilverman/booted-off-facebook-some-trump-supporters-are-bringing [https://perma.cc/4Q7W-SN4Y].

[59] Eoghan Macguire, *Banned from Facebook and Twitter, UK Far Right Turns to TikTok*, AL-JAZEERA (Apr. 16, 2020), https://www.aljazeera.com/news/2020/04/banned-facebook-twitter-uk-turns-tiktok-200416102704155.html [https://perma.cc/NAX9-L66F].

[60] Jane Li, *China's Messaging Against the Hong Kong Protests Has Found a New Outlet: Pornhub*, QUARTZ (Nov. 13, 2019), https://qz.com/1747617/chinese-users-go-to-pornhub-to-spread-hong-kong-propaganda/ [https://perma.cc/RW6D-B32Q].

[61] Levy, *supra* note 23, at 8.

[62] Cecilia D'Anastasio, *Twitch's Non-Gamers Are Finally Having Their Moment*, WIRED (Jan. 9, 2020, 4:55 PM EST), https://www.wired.com/story/twitch-non-gamers/ [https://perma.cc/A46J-A9N2].

1. Companies' Rules Are Obscure, Confusing, or Unknown to Users

For almost everyone outside the companies that make and apply them, platform rules are arcane and obscure. They are typically presented as thousands of words of fine print. Facebook's Community Standards, for example, have 26 sections–many of which include links to other documents–and a reader must navigate through each of six chapters separately.[63]

Users are unlikely to do so, to say the least. In a 2017 study, all 543 college students in a laboratory experiment clicked the "Join" button for a new social network, asserting that they had read its rules–in which they gave up their future first-born children (and all of their data) in paragraph 2.3.1.[64]

Even if one read the publicly accessible rules, they are just a rough or even misleading sketch compared with the highly detailed manuals that company staff and contractors use when regulating content.[65] For example, a ProPublica investigation revealed that although Facebook prohibits hate speech against identity groups like Black people or women, and asserted this in its public rules, its internal rules made exceptions for "subsets" of people such as female drivers or Black drivers.[66] Hate speech against subsets was acceptable on Facebook, but no one could have gleaned that from the public rules known as Community Standards.

As a result, even the most curious and diligent users cannot properly understand the vast private regulations that govern their activity online. Human-rights law should improve this problem since it requires that speech-restricting rules be precise, and accessible to those who live under them.

---

[63] *Community Standards*, *supra* note 12.

[64] David Berreby, *Click to Agree with What? No One Reads Terms of Service, Studies Confirm,* GUARDIAN (Mar. 3, 2017, 8:38 AM EST), https://theguardian.com/technology/2017/mar/03/terms-of-service-online-contracts-fine-print [https://perma.cc/SL9Y-HFRB].

[65] Some details of internet companies' internal rules have been leaked to journalists and published. *See, e.g.*, Julia Angwin & Hannes Grassegger, *Facebook's Secret Censorship Rules Protect White Men from Hate Speech But Not Black Children*, PROPUBLICA (June 28, 2017, 5:00 AM EST), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/BUY3-YDNU]; Nick Hopkins, *Revealed: Facebook's Internal Rulebook on Sex, Terrorism, and Violence,* GUARDIAN (May 21, 2017, 1:00 PM EST), https://www.theguardian.com/news/2017/may/21/revealed-facebook-internal-rulebook-sex-terrorism-violence [https://perma.cc/CX33-LAKC]. Facebook published a more detailed public version of its rules in 2018, but it was still not nearly as extensive or granular as the ones used by moderators. Josh Constine, *Facebook Reveals 25 Pages of Takedown Rules for Hate Speech and More,* TECHCRUNCH (Apr. 24, 2018, 5:02AM EST), https://techcrunch.com/2018/04/24/facebook-content-rules/ [https://perma.cc/G2N6-KWX2].

[66] Angwin & Grassegger, *supra* note 62.

2. Companies Harm People with Wrong Decisions, Both to Remove Content and to Leave It Up

When social media companies both make and enforce their rules, they must navigate a landscape of tangible harms that can result from their own action, and inaction. The gravest of these harms is mass violence, and Facebook and other companies have been excoriated for not taking down content that incites or inspires it [67] in many countries.[68] For example, in March 2018, Sri Lankan rioters burned down Muslim businesses and places of worship.[69] In an open letter, the Colombo-based Center for Policy Alternatives drew attention to Facebook posts during the riots like one in which the author "called for the 'killing of all Muslims, without sparing even a child, because they are dogs.'"[70] Six days after a user reported the post, Facebook responded that it did not violate any company rule.[71] Social media companies also err in removing content that is vital for education or for protecting human rights. For example the Syrian Archive, a project that documents human-rights abuses, has urged YouTube for years to preserve video evidence that can be vital for bringing perpetrators to justice.[72] Still YouTube has taken down over 340,000 of the Archive's 1.7 million videos.[73]

YouTube's standard appeals process has not worked well for correcting the takedowns, in part because only the person who posted a video can plead for its reinstatement. This is impossible when that person has been imprisoned or killed.[74]

Also, moderators often mistakenly remove content that is meant to educate against hatred, not express or promote it. For example, YouTube

---

[67] I coined the term "dangerous speech" for this all-too-prevalent content, and founded a research organization to find the best ways of limiting the harm such content engenders, while protecting freedom of expression. For details, see Susan Benesch et al., *Dangerous Speech: A Practical Guide*, DANGEROUS SPEECH PROJECT (Aug. 4, 2020), https://dangerousspeech.org/guide/ [https://perma.cc/AN3M-KXPU].

[68] *See, e.g.*, Alex Warofka, *An Independent Assessment of the Human Rights Impact of Facebook in Myanmar*, FACEBOOK (Nov. 5, 2018), https://about.fb.com/news/2018/11/myanmar-hria/ [https://perma.cc/XV2A-MCEU]; Stevenson, *supra* note 7.

[69] Vindu Goel, Hari Kumar & Sheera Frenkel, *In Sri Lanka, Facebook Contends with Shutdown After Mob Violence*, N.Y. TIMES (Mar. 8, 2018), https://www.nytimes.com/2018/03/08/technology/sri-lanka-facebook-shutdown.html [https://perma.cc/6H4E-3YDH].

[70] *Open Letter to Facebook: Implement Your Own Community Standards*, CTR. FOR POL'Y ALTERNATIVES, SRI LANKA (Apr. 10, 2018), https://www.cpalanka.org/open-letter-to-facebook-implement-your-own-community-standards/ [https://perma.cc/7CXJ-G6BN].

[71] *Id.*

[72] Avi Asher-Schapiro & Ban Barkawi, *'Lost Memories': War Crimes Evidence Threatened by AI Moderation*, REUTERS (June 19, 2020), https://www.reuters.com/article/us-global-socialmedia-rights-trfn/lost-memories-war-crimes-evidence-threatened-by-ai-moderation-idUSKBN23Q2TO [https://perma.cc/4LPU-F6XA].

[73] *Removal of Syrian Human Rights Content: February to April 2020*. SYRIAN ARCHIVE, https://syrianarchive.org/en/lost-found/feb-april20-takedowns [https://perma.cc/YQP4-JJZN].

[74] Hadi Al Khatib & Dia Kayyali, *YouTube Is Erasing History*, N.Y. TIMES (Oct. 23, 2019), https://www.nytimes.com/2019/10/23/opinion/syria-youtube-content-moderation.html [https://perma.cc/9868-677T].

cited its hate speech policy when removing a historical montage of the U.S. Army destroying Nazi emblems[75] and when suspending the account of history teacher Scott Allsop for hosting a channel with archival Nazi footage for his students.[76]

Adhering to international human-rights law should diminish such mistakes since it requires that any restrictions on speech be necessary and proportionate, that is, the least restrictive possible. When mistakes recur, users would have a new basis for challenging content moderation rules, and the ways in which they are enforced. The law should also reduce failures to take down dangerous content, since it requires the prohibition of advocacy of hatred that constitutes incitement to hostility, discrimination, or violence, and of propaganda for war. These provisions are described in detail below.

3. Companies Are Imposing Their Own Values

The rules and standards social media companies apply to users–and the values behind them–have their roots not in deliberative processes that take into account the norms and diversity of internet users, but rather ad hoc decision-making, mainly by a handful of U.S. lawyers.[77] While companies have made significant changes to their rules over time, they still rely upon vague and disproportionately American concepts. Some companies' leaders have realized that this is unfair. Mark Zuckerberg said in 2018: "[W]hat I would really like to do is find a way to get our policies set in a way that reflects the values of the community, so I am not the one making those decisions. . . . I think that there is likely a better process, which I haven't figured out yet." [78] Facebook has since created the Oversight Board, but its own policymaking process hasn't changed much.

Its policies are based on five values: voice, authenticity, safety, privacy, and dignity.[79] Voice–essentially freedom of expression–is "paramount," restricted only when it impinges on one of the other four val-

---

[75] Rob Beschizza, *Film of U.S. Army Destroying Nuremberg Swastikas Violates YouTube's Policy on Hate Speech*, BOINGBOING (Aug. 14, 2017, 5:33 AM EST), https://boingboing.net/2017/08/14/film-of-u-s-army-destroying-n.html [https://perma.cc/W8LZ-7Z2E].

[76] Ryan Broderick, *This History Teacher Had His Educational YouTube Channel Banned for Hosting "Hate Speech,"* BUZZFEED NEWS (June 5, 2019, 9:15 PM EST), https://www.buzzfeednews.com/article/ryanhatesthis/history-teacher-scott-allsop-youtube-channel-banned-nazi [https://perma.cc/MFU8-HUWG].

[77] Klonick, *supra* note 18, at 1621; *see* Michael Karanicolas, *Squaring the Circle Between Freedom of Expression and Platform Law*, 20 PITT. J. TECH. L. & POL'Y 175, 180-84 (2020) [https://perma.cc/FMN9-YBE6].

[78] Kara Swisher & Kurt Wagner, *Here's the Transcript of Recode's Interview with Facebook CEO Mark Zuckerberg About the Cambridge Analytica Controversy and More*," RECODE (Mar. 22, 2018), https://www.vox.com/2018/3/22/17150814/transcript-interview-facebook-mark-zuckerberg-cambridge-analytica-controversy [https://perma.cc/67KZ-UNK8].

[79] Monika Bickert, *Updating the Values that Inform Our Community Standards*, FACEBOOK (Sept. 12, 2019), https://about.fb.com/news/2019/09/updating-the-values-that-inform-our-community-standards/ [https://perma.cc/9QWT-564H].

ues.[80] In some cases, content that would otherwise be restricted is left on the platform if Facebook deems it to be sufficiently "newsworthy" or otherwise in the public interest.[81] Facebook says it "evaluate[s] the public interest value of the piece of speech against the risk of harm," through a "holistic and comprehensive" process that "account[s] for international human rights standards," [82] including Article 19 of the ICCPR.[83]

While this public commitment is somewhat reassuring, accounting for human-rights standards as part of a holistic evaluation is very different than placing them at the core of your decision-making process. Ultimately, Facebook determines "risk of harm" and "public interest" unilaterally, and imposes its decisions on billions of people.[84]

4. Shadow Censorship

Company content moderation is also used as a means for states to carry out silent and invisible censorship, by pressuring companies to take down content through undisclosed private channels or through government "Internet Referral Units" (IRUs). As Kaye reported, "Some States have pushed for social media companies and other hosts of user-generated content to monitor and take down content on their own initiative, rather than wait for law-based requests from the Government."[85] In some cases, governments claim that their requests are under color of national law, but they fail to use legal channels designed to provide due process and balance competing rights (if such protections exist at all);[86] in others, they are Mafia-style negotiations, coercing compliance from the companies by promising to ignore antitrust concerns,[87] raising the prospect of intermediary liability for user content, or outright threats to block access to the platform entirely in that country.[88]

---

[80] *Id.*

[81] *Id.*

[82] Nick Clegg, *Facebook, Elections, and Political Speech*, FACEBOOK (Sept. 24, 2019), https://about.fb.com/news/2019/09/elections-and-political-speech/ [https://perma.cc/J8AF-EUTD].

[83] 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/ [https://perma.cc/GCH3-XQ7N].

[84] *See e.g.*, Bafana Nzimande, *African Girls March Against Google and Facebook Censorship,* TIMESLIVE (Dec. 13, 2017, 3:45pm GMT+2) https://www.timeslive.co.za/news/africa/2017-12-13-african-girls-march-against-google-and-facebook-censorship/ [https://perma.cc/3B3N-Y2VK].

[85] David Kaye (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression,* U.N. Doc. A/HRC/32/38, at ¶ 45 (May 11, 2016).

[86] *Human Rights and Privatised Law Enforcement*, EUROPEAN DIGITAL RTS. (Feb. 25, 2014), https://edri.org/wp-content/uploads/2014/02/EDRi_HumanRights_and_PrivLaw_web.pdf [https://perma.cc/8P68-4T2W]; ARTICLE 19, *supra* note 25, at 16; Karanicolas, *supra* note 78, at 186.

[87] *Human Rights and Privatised Law Enforcement, supra* note 86, at 19.

[88] Klonick, *supra* note 18, at 1623.

Companies could use human-rights law as a tool to resist such pressure. It would not work magic, especially on governments with authoritarian tendencies (even those that have ratified the ICCPR), but it could help. Some governments are responsive to international-law arguments, and susceptible to being shamed for flouting them.[89] Also, companies like Facebook could use their vast power to resist improper government interference more vigorously, especially in countries where political leaders rely heavily on the use of their platforms. It would be difficult to ban (or even to slow access to) Facebook in the Philippines or Myanmar, for instance, where that platform is such a vital tool for social, economic, and political life.

Finally, grounding policies and enforcement in international human-rights standards would make it easier for users to understand what the rules are and on what rationales they are based. From there, it should be easier to influence companies to improve their rules to prevent harm, and enforce them more accurately and consistently.

III. Which International Human-Rights Law Should Companies Use?

International human rights law is found in a variety of treaties and declarations, but the UNGPs instruct businesses to focus on a few sources "at a minimum": the International Bill of Rights and the International Labor Organization's Declaration on Fundamental Principles and Rights at Work.[90] For speech regulation, the relevant documents are in the Bill of Rights, which includes the Universal Declaration of Human Rights (UDHR), the International Covenant on Civil and Political Rights (ICCPR), and the International Covenant on Economic, Social, and Cultural Rights (ICESCR).

The UDHR famously notes in its preamble that every individual and "every organ of society . . . shall strive" to promote respect for human rights and their universal and effective recognition. [91] This obviously includes companies, as the renowned human-rights scholar Louis Henkin noted.[92] The UDHR sets out the rights to freedom of expression and opinion in its Article 19, but includes no corresponding obligations.[93] It is the ICCPR that offers international law's most relevant, detailed frame-

---

[89] *See* Hilary Hurd, *How Facebook Can Use International Law in Content Moderation*, LAWFARE (Oct. 30 2019, 8:00 AM EST), https://www.lawfareblog.com/how-facebook-can-use-international-law-content-moderation [https://perma.cc/A9H5-VHWF].

[90] UNGPs, *supra* note 40, at 13 (Principle 12).

[91] UN General Assembly, *Universal Declaration of Human Rights*, Dec. 10, 1948, 217 A (III), at 1-2, https://www.un.org/en/universal-declaration-human-rights/ [hereinafter UDHR].

[92] "*Every individual* includes juridical persons. *Every individual and every organ of society* excludes no one, no company, no market, no cyberspace. The Universal Declaration applies to them all." Louis Henkin, *The Universal Declaration at 50 and the Challenge of Global Markets*, 25 BROOK. J. INT'L L. 25 (1999) [https://perma.cc/P6CS-R7MQ].

[93] UDHR, *supra* note 91, at art. 19.

work for restrictions on freedom of expression, the principal right that social media companies both facilitate and restrict. That treaty, which has been ratified by 173 of the world's 195 countries,[94] describes only two types of speech that states *must* prohibit, in Article 20: propaganda for war and advocacy of national, racial, or religious hatred that constitutes incitement to hostility, discrimination or violence.[95] Those forms of expression were familiar to the diplomats who wrote the ICCPR in the wake of World War II, and they remain all too relevant today, though some of the wording is now out of date. Still, they represent only a tiny proportion of all the content that companies restrict under their own rules.

The ICCPR's provisions on which speech *may* be restricted are found in its own Article 19, and are suited to speech regulation by social media companies in two senses. First, Article 19 requires that all restrictions be "provided by law."[96] The UN Human Rights Committee, the body charged with interpreting the ICCPR, has clarified that a norm can be considered a law.[97] Therefore Facebook's "community standards" and other companies' rules can qualify–as long as they are precise, and clearly explained to the public. A norm "must be formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly, and it must be made accessible to the public," according to the Committee.[98] This is far from the case now. Companies have been formulating their norms with increasing precision for internal use, and they have been publishing versions of their rules to the public, but those are more general and vague, as described *supra* in Section II(B)(i).

The ICCPR's flexibility also makes it apt for use by social media companies. Instead of specifying which kinds of content may be restricted, it identifies the terms on which restrictions may be imposed, and requires a transparent and balanced regulatory process. Its provisions can therefore be applied to a wide variety of rules by very different online platforms.

Another international human-rights treaty, the International Convention on the Elimination of all Forms of Racial Discrimination (ICERD),[99] is directly relevant to online content regulation, and its language conflicts with that of the ICCPR, by requiring the prohibition of different, and likely much more, speech. The ICERD's Article 4 requires

---

[94] United Nations Human Rights Office of the High Commissioner, Ratification of 18 International Human Rights Treatises, Status of Ratification Interactive Dashboard (July 28, 2020), https://indicators.ohchr.org/ [https://perma.cc/W2D4-UZHA].

[95] ICCPR, *supra* note 32 at art. 20.

[96] *Id* at art. 19(2).

[97] GC 34, *supra* note 36, at ¶ 25.

[98] *Id.*

[99] G.A. Res. 2106A (XX), at 1 (Dec. 21, 1965), https://un.org/en/development/desa/population/migration/generalassembly/docs/globalcompact/A_RES_2106(XX).pdf [https://perma.cc/NY2D-7DEE].

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."[100]

"All…ideas based on racial superiority or hatred" is surely much broader than the ICCPR's Article 20 injunction against advocacy of hatred that constitutes incitement. This apparent conflict has not been resolved or explained by the relevant authorities, though the Committee on the Elimination of Racial Discrimination, the UN body charged with interpreting that treaty, seems to have deferred to the ICCPR by directing that when speech is criminalized under Article 4 of the ICERD, "the application of criminal sanctions should be governed by principles of legality, proportionality, and necessity." [101]. In that light, and since ICERD is not part of the International Bill of Rights that the UNGPs direct companies to follow, the next section focuses on the ICCPR.

IV. How the ICCPR Articles 19 and 20 Should Be Used by Companies

The ICCPR's Articles 19 and 20 set a floor and a ceiling for restrictions on speech, and should be the core of a regulatory framework for companies.

A. Article 19

The ICCPR's Article 19 protects the right to seek and receive information of all kinds, regardless of frontiers, and through any media.[102] Article 19(3) permits restriction on speech as long as each restriction is provided by law, necessary (and proportionate), and legitimate. The last of the three prongs means that each restriction must be intended to protect at least one of five interests: national security, public order, public health, morals, and the rights and reputations of others.[103] State parties may not restrict speech for reasons *other* than those five. Also, no restriction may violate any other provision in the ICCPR, such as its ban on discrimination.[104] The three prongs of the Article 19(3) test have been explained by the Human Rights Committee, UN Special Rapporteurs, and other UN bodies. What follows are the relevant details, with my proposed application to social media companies.

---

[100] ICERD, *supra* note 35, at art. 4

[101] UN Committee on the Elimination of Racial Discrimination, *General Recommendation No. 35: Combating Racist Hate Speech,* U.N. Doc. CERD/C/GC/35, at ¶ 12 (Sept. 26, 2013), https://undocs.org/CERD/C/GC/35 [https://perma.cc/A24H-3NYD].

[102] ICCPR, *supra* note 32, at art. 19(2)

[103] GC 34, *supra* note 36, at ¶ 21.

[104] *Id.* at ¶ 26. Other protected rights include privacy, religious belief, association and peaceful assembly, education, and culture. *See* A/HRC/32/38, *supra* note 85, at ¶ 8.

1. "Provided by Law"

Though social media companies cannot make laws as such, their rules would be considered "provided by law" as long as they were "made accessible to the public" and "formulated with sufficient precision to enable an individual to regulate his or her conduct accordingly"[105] since the Human Rights Committee has clarified that norms qualify as long as these conditions are met. But most platforms' rules manifestly do not meet this standard. Twitter's CEO Jack Dorsey conceded to the U.S. Congress in 2018 that there is "a whole lot of confusion" about Twitter's rules and how they are enforced, and said, "I believe if you were to go to our rules today and sit down with a cup of coffee, you would not be able to understand them."[106] Indeed, Evelyn Mary Aswad described the vagueness of Twitter's rules in detail, in her instructive effort to apply the ICCPR's requirements to Twitter.[107] The confusion has tangible effects. As David Kaye has pointed out, it disproportionately harms the same people who are often targets of hate online: "The vagueness of hate speech and harassment policies has triggered complaints of inconsistent policy enforcement that penalizes minorities while reinforcing the status of dominant or powerful groups."[108]

To "enable an individual to regulate his or her conduct accordingly," detailed rules should be made accessible to outsiders.[109] The moderators' manuals give instructions for applying the standards to specific cases, as regulations are used to interpret statutes. Companies have long resisted releasing their manuals, saying that this would allow bad actors to "game the system"–to find ways of remaining just barely on the permissible side of a rule, or, more generally, ways of posting vicious or harmful content while avoiding takedown. This is not persuasive. First, most laws or rules draw lines between prohibited and permitted conduct–that is what law is–so conduct anywhere on the permitted side of the line is considered acceptable. If not, the line should be moved. Second, users who are determined to post harmful content and evade removal can extrapolate where the lines are, by testing the system with a variety of posts from a variety of accounts. Third, vague public rules give companies discretion to make exceptions without owning up to it.[110] Disclosure would limit this. In any case, to comply with Article 19(3), companies must publish rules in sufficient detail to allow outsiders to understand where the lines are.

---

[105] GC 34, *supra* note 36, at ¶ 25.
[106] John Eggerton, *Dorsey: Twitter Guidelines are Effectively Indecipherable,* MULTICHANNEL NEWS (Sept. 6, 2018), https://www.multichannel.com/news/dorsey-twitter-guidelines-are-effectively-indecipherable [https://perma.cc/E7EF-QVUC].
[107] Aswad, *supra* note 33, at 46-47.
[108] Kaye, *supra* note 23, at 10 (¶ 26).
[109] *See* Klonick, *supra* note 18, at 1631 (distinguishing between standards and rules).
[110] *See* Angwin & Grassegger, *supra* note 66.

2. "Necessary"

To be necessary under 19(3), as the Human Rights Committee and Kaye have explained, a restriction must be more than merely useful, reasonable or desirable.[111] It must also be strictly necessary and proportionate, meaning that the restriction is the least intrusive way of achieving the desired end, and that it is proportionate to that end.[112]

This is of special relevance to social media companies, since they have many ways of restricting content–arguably more than those available to states. Companies can of course remove content or shut down accounts. These strong measures are the most-discussed options, but as I have argued elsewhere,[113] they do not prevent harm very well since 1) by the time content is removed, harm has already been done and 2) it is easy to post the same or similar content again.

To protect freedom of expression and to prevent further harm, it is vital to try–and measure the effectiveness of–other interventions in keeping with this prong of Article 19(3). The best method is of course prevention: to persuade people not to post harmful content in the first place. Companies have begun to experiment with this a bit, by sending users messages intended to shift behavioral norms,[114] requiring users to verify their identities,[115] or temporarily suspending them, a tactic that has inevitably been nicknamed "time-outs."[116]

Companies also use other methods that are less restrictive than removing content or accounts entirely, such as downranking, or making content less accessible to fewer people on a platform.[117] On messaging

---

[111] *Id.*

[112] GC 34, *supra* note 36, at ¶ 22. *See also* David Kaye (UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression), *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, U.N. Doc A/HRC/29/32 (May 22, 2015).

[113] Susan Benesch, *Proposals for Improved Regulation of Harmful Online Content*, *in* REDUCING ONLINE HATE SPEECH: RECOMMENDATIONS FOR SOCIAL MEDIA COMPANIES AND INTERNET INTERMEDIARIES 247-306 (Yuval Shany ed., Israel Democracy Institute 2020).

[114] Susan Benesch & J. Nathan Matias, *Launching Today: New Collaborative Study to Diminish Abuse on Twitter*, MEDIUM (Apr. 6, 2018), https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc [https://perma.cc/62DB-UN6X].

[115] *Why Am I Being Asked to Upload an ID to Facebook?*, FACEBOOK, https://www.facebook.com/help/314201258613998 [https://perma.cc/P4TV-CTEF].

[116] *Twitter Rules Enforcement January to June 2019*, TWITTER, https://transparency.twitter.com/en/twitter-rules-enforcement.html [https://perma.cc/UD8T-VYQG].

[117] This is also known as "demotion" and Facebook has used it, *inter alia*, to try to prevent violence in Sri Lanka and to hide holocaust denial. See, for example, Facebook's response to a May 2020 Human Rights Impact Assessment (HRIA) on Sri Lanka, disclosing that it had demoted content that was "frequently reshared" and content from users who had violated Facebook's Community Standards before. Notably Facebook also said it seeks to respect Article 19 of the ICCPR in implementing these restrictions. *Facebook Response: Sri Lanka Human Rights Impact Assessment*, FACEBOOK (May 12, 2020), https://about.fb.com/news/2020/05/human-rights-work-in-asia/ [https://perma.cc/A5GW-H3Z3]; Kara Swisher, *Mark Zuckerberg Clarifies: "I Personally Find Holocaust Denial Deeply Offensive, and I Absolutely Didn't Intend to Defend the*

apps like WhatsApp, they have limited the number of other accounts to which a user may forward content.[118] Also, the UN Human Rights Council has encouraged states to test less restrictive measures against religious hatred, such as education and promoting interfaith dialogue.[119]

Finally, as Aswad asserted, to pass the "necessity" test companies must determine whether the measures they have chosen in fact achieve the stated aim(s). "To illustrate," she wrote, "if a company deletes posts or bans users from its platform, it needs to assess if that is helping create communities that are, for example, resilient to radicalization, knowledgeable about misinformation online, and tolerant. Similarly, a company needs to consider whether such measures cause harmful speech to fester on smaller platforms and what impact that is having on the legitimate aim." (citations omitted).[120]

This is a vital point. In complying with this requirement, companies would make powerful, public interest-protecting use of their mountains of data of many kinds, including on the details of rule-breaking. They would likely discover invaluable information about how to diminish harmful content online. Since they would be obliged to share their findings to demonstrate that they have complied with the necessity prong,[121] they would build up a valuable, shared trove of evidence-based knowledge on how to diminish harm.

3. "Legitimate," That Is, Meant to Protect One or More of Five Interests.[122]

Aswad asked whether this prong can realistically be applied to companies, since their main goal is to maximize profit. In my view, social media companies can and must be obliged to protect the public interest, just as companies in other sectors have been obliged–by law and also by public pressure–to pollute less, stop employing children, build safer cars, and so forth. Moreover, social media companies are already making decisions regarding the public interest, from trying to limit the spread of disinformation and extremist content to reducing personal online abuse–whether their motivations are altruistic or not. Profit-seeking can align with the public interest, where companies moderate content to meet the expecta-

---

Intent of Those Who Deny That," VOX: RECODE (July 18, 2018), https://www.vox.com/2018/7/18/17588116/mark-zuckerberg-clarifies-holocaust-denial-offensive [https://perma.cc/72PD-KCQ9].

[118] *More Changes to Forwarding*, WHATSAPP: BLOG (Jan. 21, 2019), https://blog.whatsapp.com/more-changes-to-forwarding [https://perma.cc/3QAG-RXUV].

[119] Human Rights Council Res. 16/18, U.N. Doc. A/HRC/RES/16/18 (Apr. 12, 2011).

[120] Aswad, *supra* note 33, at 51.

[121] Data sharing would have to be done cautiously, of course, and protect rights including privacy. Michael Karanicolas has suggested using national right to information laws and international institutions' public disclosure practices as a basis for company approaches to releasing data while protecting user privacy and trade secrets. *See* Karanicolas, *supra* note 77, at 204.

[122] ICCPR, *supra* note 31, at art. 19(3).

tions of their users and advertisers. This self-interest often overlaps with the legitimate interests discussed above, but also supports restrictions on various other types of content ostensibly not covered by Article 19 including spam or clickbait.[123]

Even when using Article 19 of the ICCPR as the basis for content restrictions, companies must not make decisions based on all five of the "legitimacy" grounds. No company may regulate based on national security, since that would be an intrusion on the sovereignty of a state. When a state itself orders a company to suppress content on legitimate national security grounds, the company must comply, but national security should come into its own decision-making only when states themselves use it improperly, trying to force companies to take down content on spurious national security grounds. This is all too common: as the Human Rights Committee has pointed out, states often try to limit freedom of expression on dubious arguments that they are protecting national security.[124] Companies are, of course, required to accept the legitimate application of national laws.

Protecting public order is also often used as a dubious basis for government repression of speech, or even for shutting down all access to the internet.[125] Companies must resist this. Since protection of public order is not constrained by national boundaries, they may regulate to prevent severe disruptions to public order, such as mass violence, of course in keeping with the "necessity" prong of Article 19(3).

Regarding the next ground, morals, Facebook and other companies already regulate on this basis. Facebook famously bans nudity, for example.[126] It purports to use the same rules for the whole world[127] though the meaning of morality varies greatly: some cultures wink at public displays of nudity, for example, and others forbid them. To allow for variations, companies that operate in many countries might adjust their rules to respect prevailing norms, as long as they still adhere to human-rights law. Companies could seek input from local, non-government councils of local advisors, as I have proposed elsewhere.[128]

The rights and reputations of others are often much easier to damage online than offline, whether through long-recognized harms like slander and libel, or newer ones like doxing and distribution of nonconsensual pornography.[129] While some remedies exist under national law,

---

[123] *See, e.g.*, Babu, Liu & Zhang, *supra* note 15.
[124] GC 34, *supra* note 36, at ¶ 30.
[125] Samuel Woodhams, *Contesting the Legality of Internet Shutdowns*, JUST SECURITY, (Oct. 1, 2019), https://www.justsecurity.org/66317/contesting-the-legality-of-internet-shutdowns/ [https://perma.cc/T8B6-S3MF].
[126] *Community Standards*, *supra* note 12, at § 14.
[127] Bickert, *supra* note 75.
[128] Benesch, *supra* note 113, at Proposal 4.
[129] For a list of such harms, see *id.,* at Part I.

they are poorly suited to the online context; regulation by social media companies can be much faster and more effective and is therefore vital. Legal remedies generally focus on holding a handful of offenders accountable after harm has been done, rather than limiting its impact in real time; they face jurisdictional issues when content crosses borders; and in many cases are simply not designed to address novel problems. However, "rights and reputations," if understood broadly, could be used as a giant umbrella to cover almost any content moderation issue, so its scope should be clarified. For example, a person might fear damage to their reputation from an online insult, but banning all insults would be much too broad an interpretation of Article 19.

Finally, social media companies can and should regulate to protect public health: both to spread useful information and to limit the spread of dangerous disinformation.[130] As the COVID-19 pandemic has demonstrated painfully well, protecting public health[131] and spreading disinformation intended to damage it[132] are both transnational enterprises, which make them well suited to regulation by social media companies.

Several companies already regulate in such ways. When Pinterest staff realized their platform was being used to spread anti-vaccine disinformation, it stopped returning search results for related terms; today such searches produce only content from public health organizations.[133] Facebook downranks such misinformation, as well as sensational health claims and products being marketed with health claims.[134] Google prohibits videos on YouTube that promote harmful remedies and cures.[135] The COVID-19 pandemic has led to stronger and more varied regulation: in addition to downranking misinformation and sensational claims, companies are fact-checking misinformation and displaying educational messag-

---

[130] Daniel Bateyko, *Dangerous Speech in an 'Infodemic'*, DANGEROUS SPEECH PROJECT (May 6, 2020), https://dangerousspeech.org/dangerous-speech-in-an-infodemic/ [https://perma.cc/SD2M-QWMH].

[131] *See, e.g.*, *Tourism Policy Responses to the Coronavirus (COVID-19)*, OECD (June 2, 2020) at Annex 1.A. https://read.oecd-ilibrary.org/view/?ref=124_124984-7uf8nm95se&title=Covid-19_Tourism_Policy_Responses [https://perma.cc/8G34-QVG2].

[132] *See, e.g.*, *Joint Communication to the European Parliament, The European Council, The Council, the European Economic and Social Committee and the Committee of the Regions, Tackling COVID-19 Disinformation–Getting the Facts Right*, at 2, COM (2020) final JOIN/2020/8 (June 10, 2020) https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020JC0008 [https://perma.cc/R7D5-HEPT].

[133] Julia Carrie Wong, *Pinterest's New Vaccine Search Will Offer Something Rare on Social Media: Facts*, GUARDIAN (Aug. 28, 2019, 12:51 PM EST), https://www.theguardian.com/society/2019/aug/28/pinterest-anti-vaccine-combat-health-misinformation [https://perma.cc/3SMB-3QFS].

[134] Louise Matsakis, *Facebook Will Crack Down on Anti-Vaccine Content*, WIRED (Mar. 7, 2019, 5:29 PM EST), https://www.wired.com/story/facebook-anti-vaccine-crack-down/ [https://perma.cc/WP5V-LWLH].

[135] YouTube Help, *Harmful or Dangerous Content Policy*, GOOGLE, https://support.google.com/youtube/answer/2801964 [https://perma.cc/89MF-P7D3].

es from health authorities,[136] removing content that contradicts public health guidance,[137] and banning advertising for products like hand sanitizer, disinfecting wipes, and test kits[138] (to reduce profiteering). [139] In sum, the public health basis for speech regulation offers companies an especially large opportunity for protecting the public interest.

B. Article 20

Article 20 describes two forms of content that must be prohibited, adhering always to the terms of Article 19. Both forms–propaganda for war and incitement to discrimination, hostility or violence–are insufficiently clear for regulation by companies, so I offer some clarification below.

Propaganda for war was such a major focus at the United Nations when the ICCPR was being drafted in the late 1940s that it became one of only two forms of speech that the treaty effectively prohibits. It has gotten short shrift in international law and UN proceedings since then,[140] perhaps since governments themselves are the source of much propaganda for war as Michael G. Kearney has noted.[141]

Still this much is clear. First, the term "war" applies only to wars of aggression in contravention of international law, not advocacy of self-determination or the right to self-determination and independence, and not civil wars.[142]

The use of the term "propaganda" instead of "incitement" indicates that the drafters of the ICCPR meant to prohibit a broader category of content than directly calling for war, or inciting a population to condone

---

[136] Guy Rosen, *An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19*, FACEBOOK (Apr. 16, 2020), https://about.fb.com/news/2020/04/covid-19-misinfo-update/ [https://perma.cc/S9GR-54PB].

[137] YouTube Help, *COVID-19 Medical Misinformation Policy*, GOOGLE, https://support.google.com/youtube/answer/9891785 [https://perma.cc/82TX-2TAH].

[138] Kang-Xing Jin, *Keeping People Safe and Informed About the Coronavirus*, FACEBOOK (Mar. 19, 2020, 2:19 PM PST) https://about.fb.com/news/2020/05/coronavirus/#ad-and-commerce-update [https://perma.cc/28VU-BFQ2].

[139] *See also* Sam Gregory, Dia Kayyali & Corin Faife, *COVID-19 Misinformation and Disinformation Responses: Sorting the Good from the Bad*, WITNESS BLOG (May 29, 2020), https://blog.witness.org/2020/05/covid-19-misinformation-response-assessment/ [https://perma.cc/5U46-2TPC] (a more comprehensive compilation and analysis of platform responses to harmful content related to COVID-19).

[140] The Human Rights Committee has produced only one comment on Article 20(1) and it is a rare model of brevity–only one page long. *See* U.N. Human Rights Committee (HRC), CCPR General Comment No. 11: Prohibition of Propaganda for War and Inciting National, Racial, or Religious Hatred (Art. 20) at 1 (July 29, 1983) https://www.ohchr.org/Documents/Issues/Opinion/CCPRGeneralCommentNo11.pdf [https://perma.cc/GFH7-F4KZ] [hereinafter GC 11].

[141] MICHAEL G. KEARNEY, THE PROHIBITION OF PROPAGANDA FOR WAR IN INTERNATIONAL LAW, at 9 (2008). This is an invaluable recent treatise on the topic.

[142] GC 11, *supra* note 140.

it.[143] They understood that turning one group of people violently against another takes time, and repeated messages.[144]

After extended debate, they "felt that a provision which was limited to prohibiting incitement to war would have little chance of securing a lasting peace and preventing future conflicts," Kearney reports.[145] Instead they chose to prohibit "the repeated and insistent expression of an opinion for the purpose of creating a climate of hatred and lack of understanding between the peoples of two or more countries, in order to bring them eventually to armed conflict." [146]

It might take courage for a social media company to restrict or remove content on the basis of Article 20(1)–and the potential benefit is enormous. As noted in Section II.B(iv), companies have more power in dealing with some governments than they have so far used.

Article 20(2) of the ICCPR requires countries to prohibit by law "[a]ny advocacy of national, racial, or religious hatred that constitutes incitement to discrimination, hostility, or violence." This unusual construction, with its use of both the terms "hatred" and "hostility" and also both the terms "advocacy" and "incitement," has inspired much confusion and debate. Article 20(2) is also somewhat outdated in naming only three bases for identity–nationality, race, and religion–not others that are found in some national laws and in the rules of social media companies, such as age, gender, sexual orientation, refugee status, or caste.

To clarify Article 20(2), then-UN High Commissioner for Human Rights Navanethem Pillay in 2012 launched a project which led to a UN document known as the Rabat Plan of Action.[147] It indicates that "'hatred' and 'hostility' refer to intense and irrational emotions of opprobrium, enmity and detestation towards the target group; the term 'advocacy' is to be understood as requiring an intention to promote hatred publicly towards the target group; and the term 'incitement' 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."[148] This presents several questions, including how social media companies, or anyone else, are to gauge when there is an imminent risk.

---

[143] KEARNEY, *supra* note 142, at 19.

[144] *See, e.g.*, Benesch et al., *supra* note 67.

[145] KEARNEY, *supra* note 142, at 131.

[146] United Nations General Assembly, *16th Session, Third Committee: Draft International Covenants on Human Rights*, U.N. Doc. A/C.3/SR.1079 (Oct. 20, 1961), at ¶ 2 https://undocs.org/A/C.3/SR.1079.

[147] *Between Free Speech and Hate Speech: The Rabat Plan of Action, a Practical Tool to Combat Incitement to Hatred*, UN OHCHR, (Feb. 21, 2013) https://www.ohchr.org/EN/NewsEvents/Pages/TheRabatPlanofAction.aspx [https://perma.cc/J3WV-LJ2W].

[148] UN High Comm'r for Human Rights, *Annual Report of the UN High Comm'r for Human Rights on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred*, app. n.5 (Rabat Plan of Action), U.N. Doc. A/HRC/22/17/Add.4 (Jan. 11, 2013) https://www.ohchr.org/Documents/Issues/Opinion/SeminarRabat/Rabat_draft_outcome.pdf [https://perma.cc/KGH7-9G7Z].

The Rabat Plan suggests a six-part threshold test for speech to be criminalized: context, speaker, intent, content and form of the speech, extent of the speech act, and likelihood, including imminence.[149] However, imminence is a poor standard for online content moderation. For example, if companies wait to respond to dangerous content until mass violence is imminent, it is usually too late to prevent it.

The Rabat plan also indicates that there must be intent to incite.[150] Intent can be very difficult to discern, especially online, and is often variable: frequently the person who originates inflammatory content intends to incite violence, but people who share it do not–or vice versa.

The Rabat test bears promise, but needs to be adapted for use in online content moderation. Another test that may be useful is the analytical framework for "dangerous speech," which I have defined as any form of human communication that can increase the risk that its audience will condone or take part in violence against another group.[151] Also, the Rabat test may be used most effectively in conjunction with sanctions other than removing content, such as restrictions on particular accounts or organizations. Finally, like other restrictions on speech, Article 20's prohibitions are subject to the legality, necessity, and legitimacy prongs in Article 19.[152]

V. Conclusion

I hope the ideas above will contribute to useful new discussion, and to further interpretation of international human-rights law for use by social media companies. There is plenty of work to be done. Without it, applying international law could make corporate content moderation more confused and ineffective, not less so.

In that light, I conclude by posing a few of many outstanding questions about how such companies might use–and comply with–the law.

- How should international human rights law contend with platforms that are end-to-end encrypted so that companies do not have access to content, as in the case of WhatsApp? Does this call for a different regulatory standard?
- What does "least restrictive means" mean in the context of platform policies, taking into account such features as encryption and objectives like user privacy?

[149] *Id.* at ¶ 29.
[150] *Id.* at 11.
[151] Benesch et al., *supra* note 67.
[152] GC 34, *supra* note 35, at 13.

- How is compliance with Article 19's legitimacy prong to be evaluated where companies regulate for several reasons, including some that are, and others that are not, specified in the treaty?
- States are permitted to enter reservations to a treaty as part of ratification. Should there be some analogous process for companies if they are otherwise unwilling to subscribe to the law?