THE FUTURE OF FREEDOM OF EXPRESSION ONLINE

EVELYN MARY ASWAD[†]

ABSTRACT

Should social media companies ban Holocaust denial from their platforms? What about conspiracy theorists that spew hate? Does good corporate citizenship mean platforms should remove offensive speech or tolerate it? The content moderation rules that companies develop to govern speech on their platforms will have significant implications for the future of freedom of expression. Given that the prospects for compelling platforms to respect users' free speech rights are bleak within the U.S. system, what can be done to protect this important right?

In June 2018, the United Nations' top expert for freedom of expression called on companies to align their speech codes with standards embodied in international human rights law, particularly the International Covenant on Civil and Political Rights (ICCPR). After the controversy over de-platforming Alex Jones in August 2018, Twitter's CEO agreed that his company should root its values in international human rights law and Facebook referenced this body of law in discussing its content moderation policies.

This is the first article to explore what companies would need to do to align the substantive restrictions in their speech codes with Article 19 of the ICCPR, which is the key international standard for protecting freedom of expression. In order to examine this issue in a concrete way, this Article assesses whether Twitter's hate speech rules would need to be modified. This Article also evaluates potential benefits of and concerns with aligning corporate speech codes with this international standard. This Article concludes it would be both feasible and desirable for companies to ground their speech codes in this standard; however, further multi-stakeholder discussions would be helpful to clarify certain issues that arise in translating international human rights law into a corporate context.

[†] The author is the Herman G. Kaiser Chair in International Law and the Director of the Center for International Business & Human Rights at the University of Oklahoma's College of Law. Previously, she served as the director of the human rights law office at the U.S. Department of State. The author wishes to thank Stanford University's Global Digital Policy Incubator for inviting her to participate in a variety of events in which issues of freedom of expression and content moderation were discussed, which contributed to her consideration of these issues. The author thanks Rebeca West for her research assistance. The views are solely those of the author.

TABLE OF CONTENTS

INTRODUCTION

In the summer days leading up to July 4th, 2018, *The Vindicator*, a small newspaper based in Liberty, Texas, decided to post on its Facebook page a few paragraphs from the Declaration of Independence.[1] Facebook blocked the tenth post because the content of those paragraphs of the Declaration violated its hate speech rules.[2] Though it did not identify what

[1] Casey Stinnett, *Facebook's Program Thinks Declaration of Independence is Hate Speech*, THE VINDICATOR (July 2, 2018, 4:46 PM), http://www.thevindicator.com/news/article_556e1014-7e41-11e8-a85e-ab264c30e973.html.

[2] *Id.* The Vindicator's tenth post contained the following language from the Declaration of Independence:

portion of the post was offensive, Facebook instructed the newspaper to remove any material that was inconsistent with its speech code.[3] The newspaper was unable to reach anyone at Facebook to appeal the decision.[4] *The Vindicator* published an article about what had happened, and this article was picked up by a number of news outlets in the United States and abroad,[5] shining a bright light on this corporate censorship like fireworks illuminating an evening sky. Despite being concerned about losing its Facebook page if future posts were also deemed unacceptable, *The Vindicator* reminded its readers that a corporation is not a governmental actor and "as such it is allowed to restrict use of its services as long as those restrictions do not violate any laws."[6] Within about a day, Facebook apologized and restored the content.[7]

Two weeks later, Facebook again made headlines about its speech code—this time for the opposite reason—when its CEO (Mark Zuckerberg) defended the company's decision to permit Holocaust denial

---

> He has abdicated Government here, by declaring us out of his Protection and waging War against us. He has plundered our seas, ravaged our Coasts, burnt our towns, and destroyed the lives of our people. He is at this time transporting large Armies of foreign Mercenaries to compleat the works of death, desolation and tyranny, already begun with circumstances of Cruelty & perfidy scarcely paralleled in the most barbarous ages, and totally unworthy the Head of a civilized nation. He has constrained our fellow Citizens taken Captive on the high Seas to bear Arms against their Country, to become the executioners of their friends and Brethren, or to fall themselves by their Hands. He has excited domestic insurrections amongst us, and has endeavoured to bring on the inhabitants of our frontiers, the merciless Indian Savages, whose known rule of warfare, is an undistinguished destruction of all ages, sexes and conditions.

Id.

[3] *Id. The Vindicator* surmised that the phrase "Indian Savages" triggered Facebook's automated detection system for hate speech and that there had been no human review of the content. *Id.* This Article uses the phrase "speech code" or "speech rules" to refer to the terms of service and other rules issued by companies that substantively regulate user-generated content on their platforms.

[4] *Id.* The newspaper did send a general feedback message to Facebook about the situation. *Id*.

[5] *See, e.g.*, Kevin Kelleher, *Facebook Reportedly Apologizes after Flagging the Declaration of Independence as Hate Speech*, FORTUNE (July 5, 2018), http://fortune.com/2018/07/05/facebook-apologizes-declaration-independence-hate-speech-racist-vindicator/; Annie Grayer, *Facebook Apologizes after Labeling Part of Declaration of Independence 'Hate Speech,'* CNN (July 5, 2018, 5:45 PM), https://www.cnn.com/2018/07/05/politics/facebook-post-hate-speech-delete-declaration-of-independence-mistake/index.html; *Facebook Finds Independence Document 'Racist,'* BBC NEWS (July 5, 2018), https://www.bbc.co.uk/news/technology-44722728.

[6] Stinnett, *supra* note 1. The first comment posted after the article argued that the First Amendment should apply to companies like Facebook. *Id.*

[7] *Id.*

posts on the platform.[8] He stated users who upload such content were not "intentionally getting it wrong."[9] Unsurprisingly, his rationale triggered a backlash of commentary given the vast proof that this atrocity happened, with many criticizing Facebook's decision to permit the hateful posts.[10] Soon thereafter, Mr. Zuckerberg clarified that he found Holocaust denials offensive and that he did not mean to defend the intent of deniers.[11] He explained his company would prevent the spread of misinformation by reducing its visibility on Facebook's News Feed, but would not prevent users from saying untrue things.[12] He did, however, note that advocacy of hatred and violence against protected groups would be removed.[13] This controversy led one commentator to say Facebook's policy "is a hodgepodge of declarations and exceptions and exceptions to the exceptions."[14] Another reflected on the controversy by musing "[i]s it

[8] Brett Molina, *Facebook CEO Mark Zuckerberg, Rebuked for Comments on Holocaust Denial, Tries to Explain*, USA TODAY (July 19, 2018, 9:22 AM), https://www.usatoday.com/story/tech/nation-now/2018/07/19/facebook-mark-zuckerberg-clarifies-comments-holocaust-deniers/799438002/; Lydia O'Connor, *Mark Zuckerberg Says Facebook Won't Remove Holocaust Denial Content*, HUFFINGTON POST (July 18, 2018, 2:53 PM), https://www.huffingtonpost.com/entry/zuckerberg-facebook-holocaust-denial_us_5b4f70f5e4b0de86f48901ea.

[9] O'Connor, *supra* note 8.

[10] *See, e.g.*, Deborah Lipstadt, *Zuckerberg's Comments Give Holocaust Deniers an Opening*, CNN (July 18, 2018, 8:43 PM), https://www.cnn.com/2018/07/18/opinions/mark-zuckerberg-facebook-holocaust-denial-lipstadt-opinion/index.html (arguing the agenda of Holocaust deniers is to "spread the very hatred that produced the Holocaust."); Molina, *supra* note 8; O'Connor, *supra* note 8.

[11] Kara Swisher, *Mark Zuckerberg Clarifies: 'I Personally Find Holocaust Denial Deeply Offensive, and I Absolutely Didn't Intend to Defend the Intent of Those Who Deny That,'* RECODE (July 18, 2018, 4:40 PM), https://www.recode.net/2018/7/18/17588116/mark-zuckerberg-clarifies-holocaust-denial-offensive.

[12] *Id.* (According to Zuckerberg, "[i]f something is spreading and is rated false by fact checkers, it would lose the vast majority of its distribution in News Feed. And of course if a post crossed [the] line into advocating for violence or hate against a particular group, it would be removed. These issues are very challenging but I believe that often the best way to fight offensive bad speech is with good speech.").

[13] *Id.* That same day, Facebook issued an official policy that would allow it to remove misinformation in the form of advocacy of incitement to violence offline. Sheera Frenkel, *Facebook to Remove Misinformation that Leads to Violence*, N.Y. TIMES (July 18, 2018), https://www.nytimes.com/2018/07/18/technology/facebook-to-remove-misinformation-that-leads-to-violence.html?rref=collection%2Fsectioncollection%2Ftechnology&action=click&contentCollection=technology&region=stream&module=stream_unit&version=latest&contentPlacement=3&pgtype=section.

[14] Farhad Manjoo, *What Stays on Facebook and What Goes? The Social Network Cannot Answer*, N.Y. TIMES (July 19, 2018), https://www.nytimes.com/2018/07/19/technology/facebook-misinformation.html.

ideal for a private company to define its own standards for speech and propagate them across the world? No. But here we are."[15]

Perhaps we should not be surprised that private actors are engaging in a parallel governance exercise alongside governments in regulating online speech.[16] In 1977, Oxford Professor Hedley Bull predicted that the international system could morph from being based on nation-states to one in which nations would share authority over their citizens with a variety of other powerful actors, including transnational corporations.[17] He called this new international order "neo-medieval" because in medieval Europe there were not nation-states but rather a variety of competing powerful actors in society that exercised various forms of governance over individuals.[18] Given that global social media companies now exercise traditional governmental functions by, among other things, enforcing their own speech codes on their platforms[19] (a process that is known somewhat euphemistically as "content moderation"), it appears that aspects of Professor Bull's neo-medieval world have materialized.[20]

---

[15] Alexis Madrigal, *Why Facebook Wants to Give You the Benefit of the Doubt*, THE ATLANTIC (July 19, 2018), https://www.theatlantic.com/technology/archive/2018/07/why-facebook-wants-to-give-you-the-benefit-of-the-doubt/565598/.

[16] Governments have also been active in regulating online speech. *See generally* SANJA KELLY, ET AL., FREEDOM HOUSE, SILENCING THE MESSENGER: COMMUNICATION APPS UNDER PRESSURE (2016), https://freedomhouse.org/sites/default/files/FOTN_2016_BOOKLET_FINAL.pdf (finding high levels of censorship by governments throughout the world for online speech that is otherwise protected under international human rights law).

[17] HEDLEY BULL, THE ANARCHICAL SOCIETY: A STUDY OF ORDER IN WORLD POLITICS 245–46, 254–266 (2d. ed. 1995). *See also* ANTHONY CLARK AREND, LEGAL RULES AND INTERNATIONAL SOCIETY 180–84 (1999) (arguing that after the Cold War the state-based system transitioned towards Professor Bull's neo-medieval system for several reasons, including the disintegration of states, the inability of states to provide for the needs of citizens, the provision of key services by transnational corporations, and the increased law making role of non-state actors).

[18] BULL, *supra* note 17, at 245.

[19] For example, Facebook had 1.47 billion daily users and 2.23 billion monthly users as of June 2018. *Company Info*, FACEBOOK NEWSROOM, https://newsroom.fb.com/company-info/ (last visited Aug. 4, 2018). The company has 7,500 content moderators who cover every time zone and 50 languages in implementing Facebook's speech code (which is found in its "Community Standards") on a worldwide basis. Ellen Silver, *Hard Questions: Who Reviews Objectionable Content on Facebook – and is the Company Doing Enough to Support Them?*, FACEBOOK NEWSROOM (July 26, 2018), https://newsroom.fb.com/news/2018/07/hard-questions-content-reviewers/ (releasing information on its content moderation because "in recent weeks, more people have been asking about where we draw the line for what's allowed on Facebook and whether our content reviewers are capable of applying these standards in a fair, consistent manner around the world.").

[20] The trajectory towards a neo-medieval world has been further accelerated by nation-states proactively outsourcing their traditional governance functions over speech to social media companies by requiring them to adjudicate whether

The rules companies develop to govern speech on their platforms will have significant implications for the future of freedom of expression and indeed democracy both in the United States and abroad. Even the U.S. Supreme Court has recognized that one of the most important places to exchange views is in cyberspace, particularly on social media.[21] But how much will it matter ten or fifteen years from now that the First Amendment (and international human rights law) protect freedom of expression, if most communication happens online and is regulated by private platforms that do not—and are not required to—adhere to such long standing substantive norms on expression?[22]

The controversies over Facebook's deletion of paragraphs from the Declaration of Independence followed by its permission of Holocaust denial posts exemplify what is now a consistent news cycle regarding private sector content moderation practices. For example, a few weeks after those controversies, major social media companies were again making headlines when they banned conspiracy theorist Alex Jones from their platforms; Twitter, however, garnered attention initially for not de-platforming him (although Twitter later suspended him and ultimately banned him permanently).[23] In recent years, there have been numerous

---

national and regional speech regulations are being violated online rather than having such issues adjudicated in courts. *See, e.g.*, Alice Cuddy, *German Law under Fire for Turning Social Media Companies into 'Overzealous Censors*,*'* EURO NEWS (Feb. 14, 2018), http://www.euronews.com/2018/02/14/german-law-under-fire-for-turning-social-media-companies-into-overzealous-censors- (discussing a recent German law that requires social media companies to decide if online speech violates the country's criminal code and to remove illegal speech or face significant penalties); Jens-Henrik Jeppesen, *First Report on the EU Hate Speech Code of Conduct Shows Need for Transparency, Judicial Oversight, and Appeals*, CTR. FOR DEMOCRACY & TECH.: BLOG (Dec. 12, 2016), https://cdt.org/blog/first-report-eu-hate-speech-code-of-conduct-shows-need-transparency-judicial-oversight-appeals/ (describing how the European Union has outsourced adjudicating its hate speech standards to social media companies).

[21] Packingham v. North Carolina, 137 S. Ct. 1730, 1735 (2017).

[22] As private sector actors, corporate speech decisions do not constitute governmental action and thus traditional sources of domestic (and international human rights) law on the permissibility of speech restrictions are not directly applicable to their actions. *See* Marvin Ammori, *The "New" New York Times: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2273–83 (2014). Furthermore, in the United States, online intermediaries are (with a few exceptions) protected from liability for third party content, giving them significant discretion to regulate speech on their platforms. *Id*. at 2284–98; Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1604–09 (2018). For a discussion of the substantial challenges to compelling platforms to respect free speech rights of users under U.S. law, *see generally* Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353 (2018).

[23] Alex Jones runs the Infowars website and has promoted a number of conspiracy theories, "such as that the Sandy Hook school shooting was a hoax and that Democrats run a global child-sex ring." Jack Nicas, *Alex Jones and Infowars Content is Removed from Apple, Facebook, and YouTube*, N.Y. TIMES (Aug. 6, 2018), https://nytimes.com/2018/08/06/technology/infowars-alefx-jones-apple-

calls for social media companies to remove various forms of offensive speech from their platforms as well as criticism that such companies delete too much speech.[24] In July 2018, Congress held a hearing to question social media companies about their content moderation practices.[25] Given

---

facebook-spotify.html. When Apple, Facebook, and YouTube removed most of Alex Jones' posts from their platforms, the tech giants thrust "themselves into a fraught debate over their role in regulating what can be said online." *Id.*; Cecelia Kang & Kate Conger, *Inside Twitter's Struggle Over What Gets Banned*, N.Y. TIMES (Aug. 10, 2018), https://www.nytimes.com/2018/08/10/technology/twitter-free-speech-infowars.html (reporting on internal deliberations at Twitter about dehumanizing speech in the wake of criticism for not banning Alex Jones); Tony Romm, *Twitter has Permanently Banned Alex Jones and Infowars*, WASH. POST (Sept. 6, 2018), https://www.washingtonpost.com/technology/2018/09/06/twitter-has-permanently-banned-alex-jonesinfowars/?utm_term=.db721d364631 (reporting on Twitter's decision to suspend and then ban Alex Jones).

[24] *See, e.g.*, Charlie Warzel, *"A Honeypot for Assholes": Inside Twitter's 10-Year Failure to Stop Harassment*, BUZZFEED NEWS (Aug. 11, 2016, 8:43 AM), https://www.buzzfeednews.com/article/charliewarzel/a-honeypot-for-assholes-inside-twitters-10-year-failure-to-s (describing Twitter's attempts to deal with abusive language given its commitment to free speech); Tracy Jan & Elizabeth Dwoskin, *Silicon Valley Escalates Its War on White Supremacy Despite Free Speech Concerns*, WASH. POST (Aug. 16, 2017), https://www.washingtonpost.com/business/economy/silicon-valley-escalates-its-war-on-white-supremacy-despite-free-speech-concerns/2017/08/16/842771b8-829b-11e7-902a-2a9f2d808496_story.html?utm_term=.8c4f8105c832 (describing platform removals of hate speech after the 2017 deadly white supremacy rally in Charlottesville); *People Don't Trust Social Media – That's a Growing Problem for Businesses*, CBS NEWS (Jun. 18, 2018, 6:45 AM), https://www.cbsnews.com/news/edelman-survey-shows-low-trust-in-social-media/ (reporting 60% of survey participants want more government regulation of social media and over 66% want companies to do more to protect users from offensive content); Nabiha Syed & Ben Smith, *A First Amendment for Social Platforms*, MEDIUM (June 2, 2016), https://medium.com/@BuzzFeed/a-first-amendment-for-social-platforms-202c0eab7054 (criticizing company speech codes as "improvised," not grounded in tradition or principle, and lacking transparency).

[25] *Facebook, Google, and Twitter: Examining the Content Filtering Practices of Social Media Giants, Hearing Before the House Judiciary Comm.*, 115th Cong. (2018), https://judiciary.house.gov/hearing/facebook-google-and-twitter-examining-the-content-filtering-practices-of-social-media-giants/ (highlighting that some representatives expressed concerns that the platforms were banning too much speech or were engaging in politically motivated content moderation while others claimed companies were not banning enough speech). In early September, the House and Senate held further hearings involving Russian misinformation online as well as content moderation. Farhad Manjoo, *What Jack Dorsey and Sheryl Sandberg Taught Congress and Vice Versa*, N.Y. TIMES (Sept. 6, 2018), https://www.nytimes.com/2018/09/06/technology/jack-dorsey-sheryl-sandberg-congress-hearings.html. Soon thereafter, then-U.S. Attorney General Jeff Sessions convened his state counterparts to discuss freedom of speech and content moderation. Brian Fung & Tony Romm, *Inside the Private Justice Department Meeting That Could Lead to New Investigations of Facebook, Google and Other Tech Giants*, WASH. POST (Sept. 25, 2018), https://www.washingtonpost.com/

significant pressure to "clean up" their platforms, some have opined that we have reached a tipping point of sorts in which social media companies are profoundly re-thinking their initial pro-free speech inclinations.[26] At the end of July, the market capitalizations of Facebook and Twitter dropped significantly, in part because of the rising costs of securing their platforms and bolstering their global content moderation.[27] In a timely and comprehensive book examining content moderation by social media companies, author Tarleton Gillespie states that "it is wholly unclear what the standards should be for content moderation."[28]

The summer of 2018 seems to mark a liminal moment in the evolution of social media speech codes that will shape the future of free expression online in our neo-medieval world. So where do we go from here? Should companies be free to set their own rules for speech on their platforms based on their economic incentives and/or own views of

---

technology/2018/09/25/inside-big-meeting-federal-state-law-enforcement-that-signaled-new-willingness-investigate-tech-giants/?utm_term=.bd73f664c69d.

[26] Julia Wong & Olivia Solon, *Does the Banning of Alex Jones Signal a New Era of Big Tech Responsibility?*, THE GUARDIAN (Aug. 10, 2018, 7:00 AM), https://www.theguardian.com/technology/2018/aug/10/alex-jones-banning-apple-facebook-youtube-twitter-free-speech ("[W]e are at an inflection point in the way internet platforms conceive of and protect public discourse for society at large."). Other commentators acknowledge a shift is occurring in how such firms approach speech but have expressed more concern about the potential consequences of private sector content moderation for freedom of expression. *See, e.g.*, Farhad Manjoo, *Tech Companies Like Facebook and Twitter are Drawing Lines. It'll be Messy*, N.Y. TIMES (July 25, 2018), https://www.nytimes.com/2018/07/25/technology/tech-companies-facebook-twitter-responsibility.html (arguing that the "absolutist ethos" of tech companies is over and expressing concerns about their power to shape global discourse through content moderation); Madrigal, *supra* note 15 ("You don't need to be a free-speech absolutist to imagine how this unprecedented, opaque, and increasingly sophisticated system [of content moderation] could have unintended consequences or be used to (intentionally or not) squelch minority viewpoints.").

[27] Peter Eavis, *The Cost of Policing Twitter and Facebook is Spooking Wall St. It Shouldn't.,* N.Y. TIMES (July 27, 2018), https://www.nytimes.com/2018/07/27/business/dealbook/facebook-twitter-wall-street.html (reporting Facebook's costs increased by 50% from 2017 to pay for, among other things, hiring hundreds of content moderators). Before their stocks tumbled, some had argued it was not sustainable for social media companies with enormous market capitalizations to have so few employees when seeking to engage in content moderation on a global scale. Henry Farrel, *The New Economy's Old Business Model is Dead,* FOREIGN POL'Y (July 13, 2018, 8:30 AM), https://foreignpolicy.com/2018/07/13/the-new-economys-old-business-model-is-dead-automation-jobs-ai-technology/ (noting pressures to regulate content online will force technology companies – which have not been big job creators relative to other major companies – to hire significantly more employees because algorithms are insufficient to deal with complex online speech issues). *See also* SCOTT GALLOWAY, THE FOUR: THE HIDDEN DNA OF AMAZON, APPLE, FACEBOOK, AND GOOGLE 266 (2017) (discussing the enormous disparity between market capitalizations and job creation by the biggest tech companies).

[28] TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET 9, 206–07 (Yale University Press 2018).

appropriate speech? Should company speech codes change based on various national laws and customs? Should governments regulate corporate content moderation?

The United Nations (UN) Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression, an independent expert appointed by UN member states who holds the top position on freedom of expression within the UN's human rights machinery, proposed a way forward during his annual report to UN member states in Geneva in June 2018. The Special Rapporteur recommended that private companies re-align their speech codes with the existing international human rights law regime.[29] He referred to social media platforms as "enigmatic regulators" that were developing an obscure type of "platform law."[30] Determining what speech is acceptable based on existing international human rights law, he argued, would give companies a universal and principled basis to engage in content moderation.[31] His recommendation to ground private sector speech codes in international standards was based on the 2011 UN Guiding Principles on Business & Human Rights, which reflect global expectations for companies to respect international human rights in their business operations.[32] In the wake of the Alex Jones controversy, Twitter's CEO tweeted that his company should root its values in international human rights law[33] and Facebook referenced human rights law in discussing its content moderation policies.[34]

Does the UN expert's recommendation to ground corporate speech codes in human rights law provide a viable (and desirable) way

---

[29] David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶¶ 3, 45, 70, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018) [hereinafter *SR Report*]. A few months earlier one of the leading international NGOs on freedom of expression made a similar call for companies to ground their speech policies in the international human rights regime. ARTICLE 19, SIDE-STEPPING RIGHTS: REGULATING SPEECH BY CONTRACT 39 (2018), https://www.article19.org/wp-content/uploads/2018/06/Regulating-speech-by-contract-WEB-v2.pdf.

[30] *SR Report*, *supra* note 29, ¶ 1.

[31] *Id.* at ¶ 42.

[32] *See id.* at ¶ 10 ("[T]he Guiding Principles on Business and Human Rights establish 'global standard[s] of expected conduct' that should apply throughout company operations and whenever they operate.").

[33] Jack Dorsey (@jack), TWITTER (Aug. 10, 2018, 9:58 AM), https://twitter.com/jack/status/1027962500438843397 [https://perma.cc/A297-PPMA].

[34] Richard Allan , *Hard Questions: Where Do We Draw the Line on Freedom of Expression*, FACEBOOK NEWSROOM (Aug. 9, 2018), https://newsroom.fb.com/news/2018/08/hard-questions-free-expression/amp/?_twitter_impression=true [https://perma.cc/Z5NP-ABEL] ("We look for guidance in documents like Article 19 of the International Covenant on Civil and Political Rights (ICCPR), which set standards for when it's appropriate to place restrictions on freedom of expression . . . . The 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.").

forward on the issue of private sector content moderation? While much of the discourse to date on content moderation has focused on increasing corporate transparency measures and improving procedural protections for users,[35] this Article focuses on the normative question of the substantive content of corporate speech codes applicable to user-generated content on their platforms. In particular, this Article seeks to unpack what the call by the UN Special Rapporteur to re-align these private sector speech rules with international human rights law would mean as a practical matter for social media companies. Part I of this Article provides background on international human rights law's protections for freedom of expression, the UN Guiding Principles on Business & Human Rights, and the recent report on content moderation by the UN's free speech expert. Part II examines what this call to re-align private sector speech rules would entail by focusing on a particular platform's hate speech code: the Twitter Rules. This examination describes aspects of Twitter's speech code that would need to change as well as raises key questions that companies, scholars, civil society, and policymakers will need to grapple with if social media companies are to respect international human rights law standards. Part III discusses potential criticisms and benefits of using international human rights law as the default for content moderation. This Article concludes that it is both feasible and desirable to ground corporate speech codes in international human rights standards while noting that the road to this desired goal, even if paved with good intentions, will be bumpy and will require further multi-stakeholder input.

I. BACKGROUND ON RELEVANT UN STANDARDS

A. International Human Rights Law & Freedom of Expression

The International Covenant on Civil and Political Rights (ICCPR) is the most relevant treaty on the topic of freedom of expression.[36] This treaty, which was opened for signature in 1966, has 172 State Parties,

---

[35] *See, e.g.*, Cindy Cohn, *Bad Facts Make Bad Law: How Platform Censorship Has Failed So Far and How to Ensure that the Response to Neo-Nazi's Doesn't Make it Worse*, 2 GEO. L. TECH. REV. 432, 447–50 (2018), https://www.georgetownlawtechreview.org/bad-facts-make-bad-law-how-platform-censorship-has-failed-so-far-and-how-to-ensure-that-the-response-to-neo-nazis-doesnt-make-it-worse/GLTR-07-2018/ (advocating for a variety of procedural protections for platform users); JOHN BERGMAYER, EVEN UNDER KIND MASTERS: A PROPOSAL TO REQUIRE THAT DOMINANT PLATFORMS ACCORD THEIR USERS DUE PROCESS, PUB. KNOWLEDGE, (May 2018), https://www.publicknowledge.org/assets/uploads/blog/Even_Under_Kind_Masters.pdf (arguing that dominant platforms should be expected to have procedures and requirements respecting users' due process); Emma Llanso, *Is Holocaust Denial Free Speech? Facebook Needs to be More Transparent*, FORTUNE: COMMENTARY (July 24, 2018), http://fortune.com/2018/07/24/facebook-mark-zuckerberg-holocaust-denial-free-speech/ (explaining that technology companies should "focus on . . . transparency, a clear appeals process, and user-empowerment tools" when removing online content).

[36] The International Covenant on Civil and Political Rights art. 19, ¶ 2, Dec. 16, 1966, S. Exec. Doc. E, 95-2 (1978), 999 U.N.T.S 171 [hereinafter ICCPR].

including the United States.[37] ICCPR Article 19 protects the right to seek and receive information of all kinds, regardless of frontiers, and through any media.[38] However, it also gives State Parties the discretion to restrict expression if they can prove that each prong of a three-part test has been met.[39] Any restrictions on speech must be

1. "provided by law" (i.e., the restriction must provide appropriate notice and must be properly promulgated) and
2. "necessary" (i.e., the speech restriction must, among other things, be the least intrusive means)
3. to achieve one of the listed public interest objectives (i.e., protection of the reputations and rights of others, national security, public order, public health or morals).[40]

These three prongs are often referred to as the legality, necessity, and legitimacy tests.[41] In addition to meeting each prong of Article 19's tripartite test, any speech restriction must also be consistent with the ICCPR's many other provisions, including its ban on discrimination.[42]

---

[37] *International Covenant on Civil and Political Rights*, UNITED NATIONS TREATY COLLECTION, https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-4&chapter=4&clang=_en (last visited Oct. 1, 2018) [hereinafter *UN Treaty Collection: ICCPR*]. The United States became a party to the ICCPR in 1992. *Id.*

[38] ICCPR, *supra* note 36, at art. 19, ¶ 2.

[39] *Id.* at art. 19, ¶ 3. The UN Human Rights Committee, the body of independent experts who are elected by the treaty's State Parties and charged with monitoring implementation of the ICCPR, has issued its recommended interpretations of Article 19 and made clear the burden of proving each prong of the tripartite test rests on the State seeking to limit speech. U.N. Human Rights Comm., General Comment No. 34, ¶ 35, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011) [hereinafter GC 34].

[40] ICCPR, *supra* note 36, at art. 19, ¶ 3. The interpretations of the tripartite test in the text above come from the Human Rights Committee's most recent guidance on Article 19. *See* GC 34, *supra* note 39, ¶ 25–34 (discussing how to interpret the ICCPR's tripartite test for restrictions on speech). The U.S. Government has interpreted the tripartite test similarly. *See* Freedom of Expression, 2011–12, DIGEST OF UNITED STATES PRACTICE IN INTERNATIONAL LAW**,** Ch.6, §L(2), at 226–27 (explaining the U.S. Government's view that "restrictions on expression must be prescribed by laws that are accessible, clear, and subject to judicial scrutiny; are necessary (e.g., the measures must be the least restrictive means for protecting the governmental interest and are compatible with democratic principles); and should be narrowly tailored to fulfill a legitimate government purpose . . . .").

[41] *See SR Report, supra* note 29, ¶ 8. (noting that restrictions on free speech must meet "the cumulative conditions of legality, necessity, and legitimacy.").

[42] The ICCPR prohibits discrimination in the implementation of treaty rights and requires State Parties to guarantee equal protection of the law without discrimination based on race, color, sex, language, religion, political or other

ICCPR Article 20(2) contains a mandatory ban on "any advocacy of national, racial, or religious hatred that constitutes incitement to violence, discrimination, or hostility."[43] This provision was highly contentious during the ICCPR negotiations; the U.S. delegation (led by Eleanor Roosevelt) and others advocated against it because it was vague and open to misuse, but the Soviet Union mustered the votes to keep it in the treaty.[44] The scope of ICCPR Article 20 remains controversial to this day. For example, a 2006 report by the UN High Commissioner on Human Rights found that governments did not agree about the meaning of the key terms in Article 20.[45] The UN even took the extraordinary measure of convening experts from around the world to propose an appropriate interpretation of this contentious sentence,[46] but this experts' process has not bridged the gap among governments with respect to Article 20's meaning. Regardless of the precise scope of Article 20, if a government seeks to restrict speech under Article 20(2), that government continues to bear the burden of surmounting the high bar set forth in Article 19's tripartite test, which significantly limits the potential reach of Article 20.[47]

---

opinion, national or social origin, property, birth, or other status. ICCPR, *supra* note 36, at arts. 2, 26.

[43] ICCPR, *supra* note 36, at art. 20, ¶ 2.

[44] For a discussion of these negotiations, see Evelyn M. Aswad, *To Ban or Not to Ban Blasphemous Videos*, 4 GEO. J. OF INT'L LAW, 1313, 1320–22 (2013). The United States became a party to the ICCPR with a reservation to Article 20 that states the article "does not authorize or require legislation or other action by the United States that would restrict the right of free speech and association protected by the Constitution and laws of the United States." *UN Treaty Collection: ICCPR*, *supra* note 37.

[45] U.N. High Commissioner for Human Rights, *Incitement to Racial and Religious Hatred and the Promotion of Tolerance*, ¶ 3, U.N. Doc. A/HRC/2/6 (Sept. 20. 2006) (finding states disagreed on the meaning of "incitement," "hatred," and "hostility").

[46] U.N. High Commissioner for Human Rights, *Rep. on the Expert Workshops on the Prohibition of Incitement to National, Racial or Religious Hatred*, ¶ 1, U.N. Doc. A/HRC/22/17/ADD. 4 (Jan. 11, 2013).

[47] *See* GC 34, *supra* note 39, ¶¶ 50–52. It should be noted that the UN Convention on the Elimination of Racial Discrimination (CERD) prohibits spreading ideas based on racial superiority or hatred as well as incitement to racial discrimination and violence. International Convention on the Elimination of All Forms of Racial Discrimination art. 4, Dec. 21, 1965, S. Exec. Doc. C. 95-2 (1978), 660 U.N.T.S. 195. Any restrictions on speech imposed under this provision also must meet ICCPR Article 19(3)'s tripartite test. *See* U.N. Comm. on the Elimination of Racial Discrimination, General Recommendation No. 35, ¶¶ 8, 12, 19, U.N. Doc. CERD/C/GC/35 (Sept. 26, 2013) (explaining that ICCPR Article 19(3)'s tripartite test of legality, necessity, and legitimacy is incorporated into this convention). The United States became a party to the CERD with the following reservation: "That the Constitution and laws of the United States contain extensive protections of individual freedom of speech, expression and association. Accordingly, the United States does not accept any obligation under this Convention, in particular under articles 4 and 7, to restrict those rights, through the adoption of legislation or any other measures, to the extent that they are protected by the Constitution and laws of the United States." *International Convention on the Elimination of*

B. The UN Guiding Principles on Business & Human Rights

1. General Background

As transnational corporate actors gained enormous power and wealth, their adverse impacts on human rights began to spark discussions at the United Nations. The debate involved whether the international human rights regime (which generally focuses on state action) could or should apply to such non-state actors. A group of independent experts tasked with making recommendations on this topic essentially proposed applying the existing human rights regime directly to companies in a 2003 document commonly referred to as "the Norms."[48] This approach was rejected by UN member states in 2004 and was generally criticized by the business community.[49] The following year, the UN Secretary General appointed a Special Representative on Human Rights and Transnational Corporations and Other Enterprises (Harvard professor John Ruggie) to try to resolve the complex issue of the appropriate relationship of international human rights law with respect to corporate actors.[50]

For six years, Professor Ruggie held numerous consultations throughout the world with stakeholders from business, civil society, indigenous groups, and UN member states.[51] He rejected the approach set forth in the Norms.[52] In 2011, he proposed to the UN Human Rights Council an alternative approach known as the Guiding Principles on Business and Human Rights (the UNGPs), which were unanimously endorsed by the Council (including the United States).[53] The U.S. Government subsequently encouraged American companies to implement the UNGPs and to treat them as a floor rather than a ceiling in their operations.[54]

All Forms of Racial Discrimination, UNITED NATIONS TREATY COLLECTION, https://treaties.un.org/Pages/ViewDetails.aspx?src=TREATY&mtdsg_no=IV-2&chapter=4&clang=_en (last visited Aug. 12, 2018).

[48] PHILIP ALSTON & RYAN GOODMAN, INTERNATIONAL HUMAN RIGHTS, 1471–72 (2d ed. 2012).

[49] *Id.* at 1477.

[50] *Id.* at 1477–78.

[51] *Id.* at 1478–79.

[52] *See id.* (Ruggie decided that the Norms were flawed because, among other things, they stated corporations were already bound by international human rights instruments, which he found had "little authoritative basis in international law.").

[53] Human Rights Council Res. 17/4, U.N. Doc. A/HRC/RES/17/4 (July 6, 2011); John Ruggie, *Rep. of the Special Representative of the Secretary-General on the issue of human rights and transnational corporations and other business enterprises to the Human Rights Council*, U.N. Doc. A/HRC/17/31 (Mar. 21, 2011) [hereinafter *UNGPs*].

[54] U.S. DEP'T OF STATE, BUREAU OF DEMOCRACY, H. R. AND LAB., U.S. GOVERNMENT APPROACH ON BUSINESS AND HUMAN RIGHTS 4 (2013), https://photos.state.gov/libraries/korea/49271/july_2013/dwoa_USG-Approach-on-Business-and-Human-Rights-updatedJune2013.pdf; U.S. GOVERNMENT, RESPONSIBLE BUSINESS CONDUCT: FIRST NATIONAL ACTION PLAN FOR THE UNITED STATES OF AMERICA 17 (2016), https://www.state.gov/e/eb/eppd/csr/naprbc/265706.htm.

The UNGPs embody the international community's expectations for how companies should act when facing human rights issues in the course of their business operations. In a nutshell, the UNGPs specify that companies should "respect" human rights, which means companies "should avoid infringing on the human rights of others and should address adverse human rights impacts with which they are involved."[55] Thus, while companies do not have all of the same human rights obligations as states do under international human rights law, corporate actors are expected to avoid adversely impacting the enjoyment of human rights and to provide remedies if rights are undermined. The UNGPs define "human rights" according to international instruments (including the ICCPR) rather than regional ones,[56] which can be less protective of human rights.[57] The UNGPs expect companies to, among other things, develop human rights policies, actively engage with external stakeholders in assessing human rights challenges, conduct due diligence to assess potential risks to human rights, and develop strategies to avoid infringing on rights.[58] Where national law conflicts with international human rights law standards, companies should seek ways to avoid infringing on human rights, but ultimately should comply with local law and address any adverse impacts.[59] The UNGPs apply to all companies regardless of size, but "the scale and complexity of the means through which enterprises meet that responsibility may vary."[60] This provides some measure of flexibility in their implementation.

2. The UNGPs in the Context of Social Media Companies

Freedom of expression represents one of the most salient human rights issues that intersects with the business operations of social media companies.[61] For social media companies to implement the UNGPs, they need to understand the scope of the right to freedom of expression under international human rights law. Additionally, social media companies should assess the risk of potential infringements on expression that occur during their business operations. Such infringements frequently happen in two ways: (1) by cooperating with governmental demands that do not meet international human rights law standards (e.g., governmental

---

[55] *UNGPs*, *supra* note 53, at Principle 11.

[56] *Id.* at Principle 12 (emphasizing that business enterprises should respect internationally recognized human rights).

[57] *See infra* notes 90–93 and accompanying text.

[58] *UNGPs*, *supra* note 53, at Principles 13–21.

[59] *Id.* at Principle 23 and accompanying commentary.

[60] *Id.* at Principle 14 and accompanying commentary ("The means through which a business enterprise meets its responsibility to respect human rights will be proportional to, among other factors, its size. Small and medium-sized enterprises may have less capacity as well as more informal processes and management structures than larger companies, so their respective policies and processes will take on different forms."). This Article focuses on the largest American social media platforms.

[61] There are other salient human rights issues that often come up in the context of social media companies (e.g., privacy), but this section focuses on expression given the overall nature and scope of this Article.

demands to remove speech critical of the head of state) and (2) by imposing their own corporate speech codes on user-generated content that restrict speech otherwise protected under international human rights law.

Although most companies do not ground their internal speech codes in international human rights law,[62] several large social media companies have already been quite active in seeking to implement the UNGPs when they face governmental demands that do not meet international human rights law standards. For example, the Global Network Initiative (GNI) involves a multi-stakeholder collaboration among companies (such as Google, Facebook, and Microsoft), investors, civil society, and academics to provide guidance about respecting freedom of expression in line with international standards.[63] GNI companies are expected to understand the scope of international freedom of expression standards and assess whether governmental demands to restrict speech comport with ICCPR Article 19 and its tripartite test (e.g., are restrictions on speech vague or not properly promulgated, are the least intrusive means used, and are regulations imposed for legitimate public interest reasons?).[64] If governmental laws or orders fail the tripartite test, GNI companies are expected to resist implementing the government's demand to the extent possible before complying with local law.[65] GNI companies may resist by, *inter alia*, initiating lawsuits in local courts and asking for the assistance of other governments or the UN's human rights machinery.[66] The GNI's assessment mechanism has consistently found that participating companies have been implementing their commitments.[67]

3. What about the Free Expression Rights of Social Media Companies?

A question that frequently arises in this context is whether expecting companies to align their internal speech codes with international human rights law violates their corporate free speech rights. While corporations have free speech rights under U.S. domestic law,[68] international human rights law protections extend only to natural persons and not to legal persons. The ICCPR provides that each State Party must respect and ensure "to all *individuals* within its territory and subject to its

---

[62] *SR Report*, *supra* note 29, ¶¶ 10, 24. Since the report was issued, Twitter and Facebook have expressed openness towards turning to international human rights law in regulating speech. *See supra* notes 33–34 and accompanying text.

[63] GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/ (last visited July 26, 2018). This initiative also covers privacy issues. *Id.*

[64] *GNI Principles*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/gni-principles/ (last visited Oct. 16, 2018).

[65] *Implementation Guidelines*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/implementation-guidelines/ (last visited July 30, 2018).

[66] *Id*.

[67] *Company Assessments*, GLOBAL NETWORK INITIATIVE, https://globalnetworkinitiative.org/company-assessments/ (last visited July 30, 2018).

[68] *See*, *e.g.*, Citizens United v. Fed. Election Comm'n., 558 U.S. 310, 342–45 (2010).

jurisdiction the rights recognized in the present Covenant."[69] The UN Human Rights Committee, the body of independent experts charged with monitoring implementation of the ICCPR and recommending interpretations of the text, has stated that only individuals (and not corporate entities) are holders of rights.[70] International law scholars have likewise taken this position.[71] Thus, requiring social media platforms to have speech codes based on international human rights law standards would not necessarily violate the speech rights of corporations under international human rights law as they do not hold such rights. That said, the UNGPs are not a legally binding framework and the U.S. government has only encouraged, not mandated, their implementation.[72] If an American platform chooses not to respect international human rights in the content and enforcement of its speech code, it would not necessarily violate international or U.S. law, but it would be acting inconsistently with the global expectations embodied in the UNGPs.

C. The UN Special Rapporteur's Report

In June 2018, the UN Special Rapporteur presented his annual report to the Human Rights Council in which he recommended "a framework for the moderation of user-generated online content that puts human rights at the very centre."[73] The report called on companies to align their content moderation policies with international human rights law and, in doing so, cited to the UNGPs.[74] In particular, the Special Rapporteur called on companies to align the substance of their speech codes with ICCPR Article 19(3)'s tripartite test of legality, necessity, and legitimacy.[75] He noted few Internet "companies apply human rights

---

[69] ICCPR, *supra* note 36, at art. 2, ¶ 1 (emphasis added). From a plain language reading, it is important that this treaty uses the word "individuals" rather than "persons," which indicates the treaty rights pertain to natural persons rather than legal persons.

[70] U.N. Human Rights Comm., General Comment No. 31, ¶ 9, U.N. Doc. CCPR/C/21/Rev.1/Add. 13 (May 26, 2004).

[71] *See, e.g.*, Thomas Burgenthal, *To Respect and To Ensure, State Obligations and Permissible Derogations*, *in* THE INT'L BILL OF RIGHTS: THE COVENANT ON CIVIL AND POLITICAL RIGHTS 73 (Louis Henkin ed. 1981) ("Juridical persons enjoy no rights under the covenant.").

[72] *See supra* note 54 and accompanying text.

[73] *SR Report*, *supra* note 29, ¶ 2.

[74] *See id. at* ¶¶ 10, 45, 70. Access Now and the Electronic Frontier Foundation, which are leading NGOs that are involved in issues of freedom of expression online, have also reaffirmed this call. *See* Access Now (@accessnow), TWITTER (Aug. 13, 2018, 5:46 PM), https://twitter.com/accessnow/status/1029167419888214016 [https://perma.cc/RK72-W927]; Electronic Frontier Foundation (@EFF), TWITTER (Aug. 13, 2018, 5:29 PM), https://twitter.com/EFF/status/1029162979453886464 [https://perma.cc/QZ4T-P2RX].

[75] *See SR Report*, *supra* note 29, ¶¶ 45–47. It should be noted that the UN expert on countering terrorism while respecting human rights has also criticized Facebook for using overly broad and vague language on terrorist content in its community guidelines and called on the company to align its speech code with international human rights law. Isa Qasim, *Exclusive: U.N. Human Rights Experts*

principles in their operations, and most that do see them as limited to how they respond to government threats and demands."[76] The report noted it would be in the companies' interests to align their internal speech codes with international human rights law because their speech codes would be grounded in universally agreed principles.[77] Rather than defending their "homegrown" versions of the appropriate parameters on worldwide speech, companies would be on firmer ground in discussions with governments (which often want them to censor too much speech) if their speech codes were aligned with international human rights protections.[78] The report also called for companies to implement a variety of improved transparency and procedural safeguards.[79]

II. ALIGNING CORPORATE SPEECH CODES WITH INTERNATIONAL STANDARDS

Currently, each social media company has its own policies about what types of speech are unacceptable on its platform.[80] As Gillespie notes in his overview of corporate speech codes, these policies often display a fundamental tension between a corporate reluctance to intervene and "a fear of not intervening,"[81] with "a range of registers on display: fussy schoolteacher, stern parent, committed fellow artist, easygoing friend."[82] In order to explore in a concrete way what re-aligning corporate speech codes to be consistent with the UNGPs and international human rights law would entail, this section examines the general approach to speech by one large social media company, Twitter. This section also analyzes Twitter's particular rules on hate speech to determine what, if anything, would need to change in such a re-alignment.[83] The analysis concludes that Twitter

---

Meet with Facebook on "Overly Broad" Definitions of Terrorist Content, JUST SECURITY (Sept. 3, 2018), https://www.justsecurity.org/60554/exclusive-u-n-rapporteur-facebook-fight-terrorist-content-risks-ensnaring/.

[76] *SR Report*, *supra* note 29, ¶ 10.

[77] *Id.* at ¶¶ 42–43, 70.

[78] *Id.* at ¶ 42.

[79] For example, the Special Rapporteur called for "radically different approaches to transparency at all stages" include sharing "case law" that shows how companies apply their speech codes. *Id.* at ¶ 71. The Special Rapporteur recommended increased disclosure of trends in decision making and called for companies to provide appeal processes and remedies for infringements on speech. *Id.* at ¶¶ 47, 58, 72. He noted that creating "social media councils" to hear complaints and rectify speech infringements could provide a scalable way forward. *Id.* at ¶ 58. He also recommended that companies subject themselves to some form of public accountability, potentially through the creation of "industry-wide accountability mechanisms[.]" *Id.* at ¶ 72.

[80] GILLESPIE, *supra* note 28, at 45–73.

[81] *Id.* at 50.

[82] *Id.* at 48.

[83] Twitter's speech code also covers a variety of topics beyond hate speech, including intellectual property issues, graphic violence and adult content, threats of violence, and the promotion of self-harm. *See The Twitter Rules*, TWITTER, https://help.twitter.com/en/rules-and-policies/twitter-rules [https://perma.cc/NXA3-2H4F]. This section is limited to consideration of Twitter's hate speech rules in order to provide a focused exposition of the analysis

would need to make substantial revisions to its rules in order to align them with international standards, and that even a good faith attempt at such re-alignment would leave some key issues open for additional discussion.

A. Twitter's General Approach to Online Speech

Twitter states that protecting users' freedom of expression is one of its core values.[84] The underpinnings of its general philosophy on speech are as follows:

> while grounded in the United States Bill of Rights and the European Convention on Human Rights, [Twitter's approach] is informed by a number of additional sources including the members of our Trust and Safety Council, relationships with advocates and activists around the globe, and by works such as [the] United Nations Principles on Business and Human Rights.[85]

Unfortunately, this foundational statement is both internally inconsistent and departs from the UNGPs. To begin with, it says Twitter's approach to freedom of expression is "grounded" in the U.S. Bill of Rights, presumably the First Amendment in particular, as well as the European Convention on Human Rights. This statement is internally inconsistent because the interpretations of free speech under the First Amendment and under the European Convention on Human Rights are often in conflict. For example, the European Court of Human Rights has upheld the banning of a blasphemous film[86] while the U.S. Supreme Court has ruled blasphemy bans unconstitutional.[87] Similarly, under U.S. law, denials of historic atrocities as well as hate speech are generally permissible as long as there is no advocacy of incitement to imminent violence or a true threat of harm.[88] The European Court of Human Rights, on the other hand, has often upheld bans on hateful speech as well as denial of historic atrocities

---

and revisions that would be needed when a social media company seeks to compare provisions in its speech code with international human rights law. This section discusses *The Twitter Rules* as they existed on August 31, 2018.

[84] *Defending and Respecting the Rights of People Using Our Service*, TWITTER, https://help.twitter.com/en/rules-and-policies/defending-and-respecting-our-users-voice [https://perma.cc/75W3-GHD7]. Although Twitter's core values and approach also encompass privacy, this section focuses on the expression aspects given the scope of this Article.

[85] *Id.*

[86] Otto-Preminger-Institute v. Austria, 295 Eur. Ct. H.R. (ser. A), ¶¶ 51–57 (1994). The European Court of Human Rights continues to highlight this 1994 case as good law in its religious freedom overview on its website. EUROPEAN COURT OF HUMAN RIGHTS, RESEARCH DIV., OVERVIEW OF THE COURT'S CASE-LAW ON FREEDOM OF RELIGION 20 (2013), http://echr.coe.int/Documents/Research_report_religion_ENG.pdf.

[87] Joseph Burstyn, Inc. v. Wilson, 343 U.S. 495, 506 (1952).

[88] Erik Bleich, *Freedom of Expression Versus Racist Hate Speech: Explaining Differences Between High Court Regulations in the USA and Europe*, 40 J. OF ETHNIC & MIGRATION STUD. 283, 283–84 (2014).

without a showing of likely lawless action in the near to midterm.[89] Given such significant divergences between American and European approaches to speech, it is unclear how Twitter's philosophy on freedom of expression can be grounded in both jurisprudential sources.

Moreover, this statement of foundational principles departs from the UNGPs, which provide that companies should seek to align their operations with *international* human rights law rather than *domestic* laws (like the U.S. Bill of Rights) or *regional* law (such as the European Human Rights Convention). Regional human rights law often departs from international law with regard to freedom of expression. For example, while the European Court of Human Rights has upheld blasphemy bans,[90] the UN Human Rights Committee, which recommends interpretations of the ICCPR, has generally condemned bans on blasphemous speech.[91] Similarly, the European Court of Human Rights has upheld criminal bans on speech that denies historic atrocities,[92] whereas the UN Human Rights Committee has disapproved of such censorship.[93] In sum, *regional* human rights instruments (and monitoring bodies) are not *international* human rights instruments (and monitoring bodies). Thus, the scope of protection afforded under each may differ. As a statement of global expectations, the UNGPs are properly pinned to international instruments and not regional instruments, unlike Twitter's general philosophy on speech.

It should also be noted that, by selectively highlighting one region's human rights convention (i.e. Europe), Twitter opens itself up to claims from countries in other regions that their own regional human rights instruments should be used to evaluate speech uploaded or viewed in their parts of the world. Those regional instruments can also depart from the ICCPR and provide fewer protections. For example, the Human Rights Declaration of the Association of South East Asian Nations[94] (ASEAN) limits rights, including freedom of expression, in a variety of ways that are inconsistent with international standards.[95] The Organization of Islamic

---

[89] *Id*. *See also* Noah Feldman, *Free Speech in Europe Isn't What Americans Think*, BLOOMBERG: VIEW (Mar. 19, 2017, 9:33 AM), https://www.bloomberg.com/view/articles/2017-03-19/free-speech-in-europe-isn-t-what-americans-think.

[90] *Otto-Preminger-Institute*, 295 Eur. Ct. H.R. (ser. A), ¶¶ 51–57.

[91] GC 34, *supra* note 39, ¶ 48 ("Prohibitions of displays of lack of respect for a religion or other belief system, including blasphemy laws, are incompatible with the Covenant, except in the specific circumstances envisaged in article 20, paragraph 2, of the Covenant.").

[92] Bleich, *supra* note 88, at 283–84.

[93] GC 34, *supra* note 39, ¶ 49 ("Laws that penalize the expression of opinions about historical facts are incompatible with the obligations that the Covenant imposes on States parties in relation to the respect for freedom of opinion and expression. The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events.").

[94] *ASEAN Human Rights Declaration*, ASSOCIATION OF SOUTHEAST ASIAN STATES (Nov. 19, 2012), http://asean.org/asean-human-rights-declaration/.

[95] The ASEAN Declaration's inappropriate limitations on rights include "the use of the concept of 'cultural relativism' to suggest that rights in the [Universal Declaration on Human Rights] do not apply everywhere; stipulating that domestic

Cooperation, which is comprised of 57 nations,[96] has formed a human rights system. This system is based in part on the Cairo Declaration on Human Rights in Islam, which explicitly limits free speech according to Shariah norms.[97] What basis does Twitter have for favoring (or applying) Europe's regional approach to human rights in its global operations over other regions' human rights instruments? It is only by citing to universal standards embodied in international human rights law that Twitter can claim to ground its worldwide rules in a fair manner.

If Twitter were to heed the UN Special Rapporteur's call to act consistently with the UNGPs, the company would need to revise the general philosophy underlying its speech code by making at least two key changes. First, rather than highlighting any particular domestic laws or regional human rights instruments, the philosophical statement should reference a commitment to aligning its approach to speech with international human rights law and ICCPR Article 19 in particular. Second, Twitter's approach should not be "informed" by the UNGPs, but rather it should clearly commit to "implementing" the UNGPs. Such fundamental revisions in its basic philosophy would result in a shift with respect to the substance and execution of its speech code and warrant appropriate training to mainstream a new approach grounded in international human rights law.

B. Twitter's Approach to Hate Speech

Twitter's hate speech provisions appear under the "hateful conduct" and "hateful imagery and display names" headings of its speech code.[98] With respect to hateful conduct, users may not promote "violence against, threaten, or harass other people on the basis of race, ethnicity, national origin, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease."[99] Prohibited hate speech is further defined as, among other things, speech that harasses by wishing for harm of individuals or groups, inciting fear of a protected group, and repeating content that degrades someone.[100] Decisions about whether

---

laws can trump universal human rights; incomplete descriptions of rights that are memorialized elsewhere; introducing novel limits to rights; and language that could be read to suggest that individual rights are subject to group veto." Press Release, U.S. Dep't of State, ASEAN Declaration on Human Rights Press Statement (Nov. 20, 2012), https://2009-2017.state.gov/r/pa/prs/ps/2012/11/200915.htm.

[96] *Member States*, ORGANISATION OF ISLAMIC COOPERATION, https://www.oic-oci.org/states/?lan=en (last visited Aug. 10, 2018).

[97] The Organisation of the Islamic Cooperation, *The Cairo Declaration on Human Rights in Islam* art. 22, Aug. 5, 1990, Annex to Res. No. 49/19-P, *available at* https://www.oic-iphrc.org/en/data/docs/legal_instruments/OIC_HRRIT/571230.pdf.

[98] TWITTER, *supra* note 83. Other forms of abusive speech are also covered by Twitter's rules, but the focus of this Article is on the company's hateful conduct, hateful imagery, and hateful display policies.

[99] *Id.*

[100] *Hateful Conduct Policy*, TWITTER: HELP CENTER, https://help.twitter.com/en/rules-and-policies/hateful-conduct-policy [https://perma.cc/LU63-AVTP].

violations have occurred are made by referring to context, including discussions with aggrieved persons.[101] Users are also prohibited from using "hateful images or symbols in [their] profile image or profile header" as well as using their "username, display name, or profile bio to engage in abusive behavior, such as targeted harassment or expressing hate towards a person, group, or protected category."[102]

There are a variety of consequences for violations. For example, Twitter can make tweets less visible in several ways, including with respect to search results.[103] Twitter can also prevent rule violators from tweeting again until they delete tweets that cross the line and can hide the tweets until they are deleted.[104] If a profile is non-compliant, Twitter can make it unavailable until it is changed.[105] An account can also be placed in "read only" mode, limiting the violator's ability to tweet, retweet, or like content.[106] Twitter's most severe penalty is permanent account suspension, which removes the account from view (and violators are prohibited from creating new Twitter accounts).[107] The type of reprimand is based on a variety of factors, including the severity of the violation, the user's track record of behavior on Twitter, and whether the topic may be of legitimate public interest.[108]

1. Is Twitter's Hate Speech Ban Vague?

As discussed in Part I, for any restriction on speech to be valid under the ICCPR, the restriction must be (1) "provided by law" (e.g., not vague) and (2) "necessary" (e.g., the least intrusive means) (3) to achieve a legitimate aim.[109] If we apply the ICCPR's tripartite test in the context of Twitter's hate speech rules, it is clear that aspects of Twitter's rules would need revision, particularly with respect to the requirement that speech restrictions not be vague (as many terms in Twitter's hate speech ban are unclear). For example, what is the scope of speech that constitutes "expressing hate" towards someone or a group? What range of speech "degrades" someone? Which images would meet the "hateful" threshold? The UN Special Rapporteur found, as a general matter, that "[c]ompany policies on hate, harassment, and abuse also do not clearly indicate what constitutes an offence," and he highlighted, in particular, Twitter's prohibition on speech that "incites fear about a protected group" as

---

[101] *Id.*
[102] TWITTER, *supra* note 83.
[103] *Our Range of Enforcement Options*, TWITTER: HELP CENTER, https://help.twitter.com/en/rules-and-policies/enforcement-options [https://perma.cc/F6VD-E7X3].
[104] *Id.*
[105] *Id.*
[106] *Id.*
[107] *Id.*
[108] *Our Approach to Policy Development and Enforcement Philosophy*, TWITTER, https://help.twitter.com/en/rules-and-policies/enforcement-philosophy [https://perma.cc/89KT-DJ9A].
[109] *See supra* notes 36–47 and accompanying text.

subjective and vague.[110] Twitter would need to revise its hate speech rules to address a variety of vagueness issues to pass the "legality" prong of the ICCPR's tripartite test.[111]

2. Do Twitter's Hate Speech Rules Constitute the "Least Intrusive Means?"

While the UN Special Rapporteur's report provided some tangible guidance on what it would look like for companies to respect the first prong of ICCPR Article 19(3)'s tripartite test,[112] his commentary does not address details of how a company would apply the second or third prongs of Article 19(3) in its operations.[113] A company may consider approaching Article 19(3)'s second prong (i.e., the "necessity" or "least intrusive means" test) by drawing on lessons learned from governments' experiences. This section reviews some of those experiences and proposes a company would need, at a minimum, to publicly commit to three steps to act consistently with the "necessity" prong of the tripartite test. First, a company should evaluate the means at its disposal to achieve a legitimate aim without infringing on speech. Second, in assessing various options that infringe on speech, a company should select the option that reflects the least intrusion on speech interests. Third, the company should periodically assess whether the selected measure helps to achieve the legitimate aim or not. Each step involves an analysis that differs from—but can be usefully informed by—how governmental actors are expected to analyze these issues.

Turning to the first step, what are the types of options available to companies to achieve legitimate aims that do not involve infringing on

---

[110] *SR Report*, *supra* note 29, ¶ 26. The Special Rapporteur is not alone in his concern that company speech codes are vague. During a Congressional hearing in September 2018, Twitter's CEO acknowledged "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." Manjoo, *supra* note 25. (Perhaps the best example of the vagueness issues with the Twitter Rules is that the company initially felt Alex Jones had not violated its speech code, then determined he merited a temporary suspension, and then de-platformed him. *See supra* note 23 and *infra* note 141.) It should also be noted that, under the ICCPR, any speech restrictions must also comply with the treaty's other protections, including its ban on discrimination. Article 26 of the ICCPR provides equal protection for a wide array of groups. ICCPR, *supra* note 36, at art. 26. Twitter's list of protected groups is not as broad as the ICCPR's list of groups because the company's list does not, for example, cover political groups. *See supra* note 99 and accompanying text.

[111] Corporations have an incentive to keep their speech codes vague because it helps them take the position they are correct in whatever enforcement action they choose to implement. That said, if a company pledges to respect Article 19(3) in its speech code, it would be possible for civil society and users to assess if the company has overcome its inclinations or has maintained vague speech codes.

[112] *SR Report*, *supra* note 29, ¶¶ 26–27, 46 (discussing vagueness problems with company speech codes).

[113] *Id.* at ¶¶ 28, 47 (calling for increased transparency when discussing the second prong of the tripartite test rather than providing granular guidance for implementation).

speech, including deleting speech or blocking speakers from their platforms? To answer this question in the context of corporate actors, it is instructive to examine the toolkit that the international community has agreed governments should use to fight religious hatred and intolerance. For over ten years at the United Nations, countries fought about whether it was necessary to ban blasphemy, speech that embodies religious hatred, or speech that otherwise greatly offends religious sensibilities to promote religious tolerance.[114] Some nations argued it was necessary to ban such speech not only to promote tolerance, but also for individuals to feel comfortable to practice their religious beliefs and to feel safe in society.

In 2011, the international community ultimately determined in UN Human Rights Council Resolution 16/18 that governments have a host of options short of broad bans on speech to end religious hatred and promote tolerance.[115] These options include promoting relevant educational initiatives and inter-faith dialogues, training government employees in effective outreach strategies to vulnerable groups, encouraging government officials to speak out against intolerance, and robustly implementing discrimination and hate crimes laws (i.e., punishing discriminatory behavior as a way of preventing potential harmful impacts of intolerant speech).[116] This resolution only calls for a ban on speech when there is incitement to imminent violence, which reflects the U.S. constitutional standard for banning speech that incites harm.[117] This set of proactive good governance actions, short of broad speech bans, is often referred to as the "16/18 consensus toolkit."[118] Under this rubric, it would be inappropriate for a government to resort to banning offensive speech to promote religious tolerance if the government had not even tried to engage in the basic good governance measures set forth in the 16/18 toolkit. In other words, resorting to speech bans without engaging in good governance measures would not constitute the "least intrusive means" to achieving religious tolerance and public order.

When considering what options companies should consider before infringing on speech in situations involving, among other things, online hatred and intolerance, it is helpful to keep this 16/18 toolkit in mind. Like governments, companies can also speak out on issues, educate users, and promote dialogue on contentious issues. It seems that companies have already been implementing some activities similar to those in the 16/18 toolkit to help them tackle several pressing issues. For example, Facebook has been funding a variety of dialogue and counter-narrative approaches to combatting hate and violent extremism.[119] Google has been funding

---

[114] Aswad, *supra* note 44, at 1323.
[115] Human Rights Council Res. 16/18, U.N. Doc. A/HRC/RES/16/18 (Apr. 12, 2011) [hereinafter Council Res. 16/18].
[116] *Id.* at ¶¶ 5–6.
[117] Aswad, *supra* note 44, at 1325.
[118] *Id.* at 1328.
[119] *See, e.g.*, Jeremy Kahn, *How Facebook Can Fight the Hate*, BLOOMBERG BUSINESSWEEK (May 25, 2017, 4:00 AM), https://www.bloomberg.com/news/features/2017-05-25/how-facebook-can-fight-the-hate (discussing specific

educational initiatives on media literacy to help combat misinformation online.[120] Like governments, companies can and should be creative and proactive in developing actions inspired by the 16/18 toolkit that can help resolve issues without infringing on speech on their platforms. As noted previously, the UNGPs provide a measure of flexibility in their implementation based on the size and resources of a company, which will be of particular relevance for smaller social media companies in developing appropriate toolkits.[121]

After implementing available "good governance" measures, companies should consider the second step noted above to act consistently with Article 19(3)'s necessity test. When a company must resort to infringing on speech, it should carefully develop a continuum of options for dealing with problematic speech and commit publicly to selecting the least intrusive means to resolve the problem. In the context of governments, the least intrusive means test often involves, for example, selecting civil rather than criminal sanctions for harmful speech.[122] For private platforms, there are a range of actions to be considered. For example, a company could give its users a means to opt out of offensive material.[123] Another option could be that a company avoids giving problematic posts a circulation boost, but does not delete them or affect its users' ability to circulate the posts.[124] A company could also lower the ranking of problematic posts in search results or otherwise decrease their visibility.[125] Although options involving de-emphasizing posts would not

---

counter-narrative measures to combat extremism online that are funded by Facebook).

[120] *See* Kevin Roose, *Google Pledges $300 Million to Clean Up False News*, N.Y. TIMES (Mar. 20, 2018), https://www.nytimes.com/2018/03/20/business/media/google-false-news.html (reporting Google promised $10 million to help teenagers identify misinformation).

[121] *See supra* text accompanying note 60.

[122] For instance, the UN Human Rights Committee has advised State Parties to the ICCPR to avoid criminal sanctions in the context of defamation suits. GC 34, *supra* note 39, ¶ 47.

[123] *See, e.g.*, Cohn, *supra* note 35, at 451 (arguing Facebook gives users the ability to choose the types of ads they prefer and could develop "a similar system" so users could avoid offensive content rather than Facebook banning the content); Llansó, *supra* note 35 (proposing alternatives to banning speech such as "involving members of the site's community in administering and moderating subsections based on those sections' own norms and policies, or allowing individual users to set their own filters and rules for what they can see and share on the site.").

[124] Facebook, for instance, has stated it will not remove Holocaust denial posts, but will not give them a circulation boost in its News Feed. Swisher, *supra* note 11.

[125] After President Trump accused Twitter of "shadow banning" Republican tweets, Twitter released a statement explaining it does not shadow ban (which it defined as making tweets "undiscoverable to everyone except to the person who posted it") and explained it ranks search results by boosting those tweets that are relevant to users and popular and de-emphasizing tweets "from bad-faith actors who intend to manipulate or divide the conversation" in order to promote a "healthy conversation." Vijaya Gadde & Kayvon Beykpour, *Setting the Record*

delete them from the platforms, this analysis treats them as an infringement on speech as such techniques could have the effect of essentially burying posts.[126] Where speech must be banned, geo-blocking (i.e., restricting access to Internet content based on location) a particular post from view in the particular country could be considered (rather than removing the information from the platform).[127] A more intrusive infringement on speech on this continuum would be to delete a post but to allow the speaker to continue to speak on the platform.[128] Warnings could be issued to a user who repeatedly violates a company's speech code before taking more severe measures. The most extreme end of the continuum may be blocking a user's account in egregious situations.[129] In

---

Straight on Shadow Banning, TWITTER (July 26, 2018), https://blog.twitter.com/official/en_us/topics/company/2018/Setting-the-record-straight-on-shadow-banning.html. Twitter's speech code also specifies that its range of enforcement actions include making a tweet less visible based on various factors, including the "quality of the content." *Our Range of Enforcement Actions*, *supra* note 103. Google has also redirected search results to help counter violent extremism. Kahn, *supra* note 119.

[126] *See* Tessa Lyons, *Replacing Disputed Flags with Related Articles*, FACEBOOK NEWSROOM (Dec. 20, 2017), https://newsroom.fb.com/news/2017/12/news-feed-fyi-updates-in-our-fight-against-misinformation/ ("Demoting false news (as identified by fact-checkers) is one of our best weapons because demoted articles typically lose 80 percent of their traffic."). Very little is known about how companies engage in such practices that de-emphasize information on their platforms. The need for greater clarity and transparency about how companies affect discourse through ranking information on their platforms continues to be a crucial aspect of understanding how they regulate – and affect – speech and therefore of assessing the extent to which such measures infringe on speech relative to other measures.

[127] For example, it may be the case that geo-blocking advocacy to incitement to imminent violence in a particular country could help avoid an atrocity in that country, but allowing those outside the country to view the speech could help formulate responses by the international community, including gathering evidence for accountability purposes.

[128] Given the global scale at which many social media companies operate, there may be a temptation to rely too much on automated methods to delete speech that violate speech codes (as appears to have occurred when Facebook deleted the post containing parts of the Declaration of Independence). As noted in a report by the Center for Democracy and Technology, it is wrong to "assume that automated technology can accomplish on a large scale the kind of nuanced analysis that humans can accomplish on a small scale." CTR. FOR DEMOCRACY & TECH., MIXED MESSAGES? THE LIMITS OF AUTOMATED SOC. MEDIA CONTENT ANALYSIS 3 (Nov. 2017), https://cdt.org/files/2017/11/Mixed-Messages-Paper.pdf. Relying solely on automated technology is likely to delete too much speech and thus be inconsistent with the least intrusive means test.

[129] For example, Twitter's rules note that the company's most severe enforcement action is permanent account suspension. *Our Range of Enforcement Options*, *supra* note 103. Getting kicked off a major platform has been referred to as the "death penalty" in the digital world. *See* Will Sommer, *YouTube Bans Infowars' Alex Jones from Spewing Hate Speech*, DAILY BEAST (Aug. 6, 2018, 12:06 PM), https://www.thedailybeast.com/youtube-bans-infowarss-alex-jones-for-spewing-

sum, a variety of corporate options exist that infringe on speech to varying degrees, and a company bears the burden of proving it has selected the least intrusive means in acting consistently with ICCPR Article 19(3).[130]

Finally, the third step a company should undertake regarding the "necessity" test is diligently monitoring whether the measure it has selected is helping to further a legitimate aim. To illustrate, 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.[131] 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.[132] A company should assess whether its selected measures have negative unintended consequences[133] that may outweigh the desired benefits and whether such measures unproductively

---

hate-speech ("In recent weeks, Facebook, Apple, and Spotify had banned Infowars, but YouTube had seemed reluctant to impose the death penalty.")

[130] This section does not comprise a comprehensive listing of options. For additional options, see Mike Masnick, *Platforms, Speech, and Truth: Policy, Policing and Impossible Choices*, TECHDIRT (Aug. 9, 2018, 9:42 AM), https://www.techdirt.com/articles/20180808/17090940397/platforms-speech-truth-policy-policing-impossible-choices.shtml. Another potential option to be considered could involve time-limited content blocking when there are substantial risks about immediate violence that would not trigger the same concerns after a particular situation is diffused.

[131] Companies have shown an understanding of monitoring whether their selected approaches work or not. For instance, when Facebook found that "flagging" misinformation was not helpful in combatting misinformation, it switched to circulating related articles to give context to misinformation to better combat it. Lyons, *supra* note 126.

[132] *See* Joanna Plucinska, *Hate Speech Thrives Underground*, POLITICO (Feb. 7, 2018, 12:12 PM), https://www.politico.eu/article/hate-speech-and-terrorist-content-proliferate-on-web-beyond-eu-reach-experts/ (reporting that "with increased scrutiny on mainstream sites, alt-right and terrorist sympathizers are flocking to niche platforms where illegal content is shared freely, security experts and anti-extremism activists say."); *see also* Jessica Schulberg et al., *The Neo-Nazis Are Back Online*, HUFF. POST (Oct. 3, 2017, 9:43 PM), https://www.huffingtonpost.com/entry/nazis-are-back-online_us_59d40719e4b06226e3f46941 (describing how Stormfront, a neo-Nazi internet forum, was able to transfer its domain name from one domain registrar to another after being shut down).

[133] *See, e.g.*, Rob Price, *YouTube's Crackdown on Extremist Content and ISIS is also Hurting Researchers and Journalists*, BUSINESS INSIDER (Aug. 14, 2017, 7:30 AM), https://www.businessinsider.com/youtube-crackdown-terrorist-extremist-isis-content-hurting-journalists-researchers-2017-8?r=UK&IR=T; J.M. BERGER & JONATHON MORGAN, THE ISIS TWITTER CENSUS: DEFINING AND DESCRIBING THE POPULATION OF ISIS SUPPORTERS ON TWITTER 54–58 (The Brookings Project on U.S. Relations with the Islamic World, Mar. 2015), *available at* https://www.brookings.edu/wp-content/uploads/2016/06/isis_twitter_census_berger_morgan.pdf (noting account suspensions could result in potential loss of key information for law enforcement and terror networks could turn insular, reducing de-radicalizing influences).

raise the profile of harmful speech and speakers.[134] If the selected measures are infringing on speech without furthering the legitimate aim, the company needs to reconsider its options.

In aligning its speech code to the "necessity" prong of ICCPR Article 19's tripartite test, Twitter would need to make some revisions to its existing speech code. First, it should commit publicly to investigating (and investing in) "good governance" measures that do not infringe on speech to the extent possible. Second, while Twitter is to be commended for setting forth a broad range of enforcement options, it should commit publicly to ensuring its selected response is calibrated to constitute the least intrusive means.[135] Third, Twitter should also commit to monitor closely whether the measures it undertakes to promote legitimate aims are working. If a measure that infringes on speech is not helping to achieve a legitimate aim, Twitter should revise its approach accordingly.

3. Is Twitter's Hate Speech Ban Imposed for a Legitimate Aim?

The third prong of ICCPR Article 19(3)'s tripartite test requires that any speech restriction be imposed for one of the following legitimate aims that benefit the public interest: respect of the rights or reputations of others; or the protection of national security, public order, public health, or morals.[136] Under this "legitimacy" test, it would be an invalid reason and a violation of the ICCPR for a government to impose a speech ban to end criticism of the head of state (even if administered under the pretext of "public order").[137] On the other hand, it would be legitimate for a government to invoke a public order rationale if the true motive for a speech ban were to stop advocacy likely to result in imminent violence against a vulnerable minority.[138] In sum, this third prong requires the government to identify in good faith one of the legitimate public interest reasons for restricting speech.

In translating Article 19(3)'s legitimacy prong from the governmental context to the corporate context, a threshold question arises: can we expect corporations to make such public interest determinations

---

[134] *See* Masnick, *supra* note 130 ("[De-platforming] someone from these platforms often has the opposite impact of what was intended. Depending on the situation, it might not quite be a 'Streisand Effect' situation, but it does create a martyr situation, which supporters will automatically use to double down on their belief that they're in the right position, and people are trying to 'suppress the truth' or whatever. Also, sometimes it's useful to have 'bad' speech out in the open, where people can track it, understand it... and maybe even counter it. Indeed, often hiding that bad speech not only lets it fester, but dulls our ability to counter it, respond to it and understand who is spreading such info (and how widely).").

[135] As noted previously, further transparency in terms of how these enforcement mechanisms operate in practice is essential to allow civil society, academics, and others to assess whether least intrusive means are being selected. *See supra* text accompanying note 126.

[136] ICCPR, *supra* note 36, at art. 19, ¶ 3.

[137] *See* GC 34, *supra* note 39, ¶ 38 ("Moreover, 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.").

[138] Aswad, *supra* note 44, at 1322.

when restricting speech? As Professor Klonick has observed, the main reason companies remove offensive speech is "the threat that allowing such material poses to potential profits based in advertising revenue."[139] Companies are essentially seeking to moderate content on their platforms in order to meet user expectations so they can maximize profits.[140] In our neo-medieval world, are advertisers essentially the ultimate judges when it comes to the content of speech codes? Can we expect corporations to refrain from restricting speech at the expense of their bottom lines? Does gauging the temperature of most users in determining the scope and application of speech codes boil down to rule by the majority at the expense of the minority in our neo-medieval world?[141]

The legitimacy prong of ICCPR Article 19's tripartite test seems to pose the thorniest questions when translating its requirements from the context of governments to that of corporate actors. There are two main options worth considering: exempt companies from this prong or hold them to the public interest assessments contained in Article 19(3).[142]

---

[139] Klonick, *supra* note 22, at 1627.

[140] *Id.* at 1627 ("If a platform creates a site that matches users' expectations, users will spend more time on the site and advertising revenue will increase. Take down too much content and you lose not only the opportunity for interaction, but also the potential trust of users. Likewise, keeping up all content on a site risks making users uncomfortable and losing page views and revenues."). *See also* GILLESPIE, *supra* note 28, at 17 ("[F]rom an economic perspective, all this talk of protecting speech and community glosses over what in the end matters to platforms more: keeping as many people on the site spending as much time as possible, interacting as much as possible."). Companies may also modify their speech codes (or how robustly they enforce them) in response to pressure from governments that are less protective of speech and then apply those standards worldwide in their terms of service. *See* Danielle Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1041–50 (2018) (describing how U.S. platforms altered their content moderation on extremism and hate speech in response to threats of European regulation).

[141] In the news coverage of prominent platforms banning Alex Jones, some raised the issue of whether public pressure had triggered the de-platforming. *See, e.g.*, Nicas, *supra* note 23. Twitter's CEO initially defended his company's decision not to ban Mr. Jones by saying "[i]f we succumb and simply react to outside pressure, rather than straightforward principles we enforce (and evolve) impartially regardless of political viewpoints, we become a service that's constructed by our personal views that can swing in any direction. That's not us." Jack Dorsey (@jack), TWITTER (Aug. 7, 2018, 5:11 PM), https://twitter.com/jack/status/1026984247750316033 [https://perma.cc/8QWN-A25M]. Twitter later interpreted its rules to merit a suspension (and ultimately a permanent ban) of Alex Jones. Romm, *supra* note 23.

[142] Perhaps a third option could be to allow companies to consider their economic incentives along with public interest reasons when restricting speech, but implementation of such an option in practice would likely be dominated by economic motivations, thereby risking that public interest rationales become mere pretexts for decisions based on revenue. Rather than incentivizing the use of public interest rationales as pretexts, it is better to be transparent that the legitimacy prong is removed for companies (i.e., the first option) or to truly hold

Under the first option, companies would be exempted from having to justify their restrictions on speech based on public interest considerations. Because of their nature as entities designed to maximize shareholder profits, this option assumes that companies cannot be expected to engage in public interest determinations that conflict with their earning potential. This option suggests that companies would cite to public interest determinations as pretexts for revenue-driven outcomes. If we proceed on this basis, companies would still be expected to respect the first two prongs of the tripartite test, i.e., companies should make sure their speech codes are not vague and select the least intrusive enforcement measure for violations. If companies publicly commit to implementing the legality and necessity tests of ICCPR Article 19(3), society would indeed be in a better place in terms of protecting freedom of expression in our neo-medieval world than what is currently happening.

However, this option (which no longer requires speech restrictions be linked to the public interest) means that as long as companies adhere to their own rules that are not vague and only infringe on speech in the least intrusive way, they could restrict speech for any reason at all, including maximizing their revenues or promoting their favored policies. If the normative goal for our neo-medieval world is to develop a path that continues protecting freedom of expression despite the enormous power of private platforms over speech, removing the third prong of the tripartite test is unappealing. Allowing advertising dollars to ultimately decide the contours of appropriate speech for platforms does not present a particularly attractive future. And neither is a future in which norms are established by the whims of the majority of users (or of the most vociferous users). In such systems, minority and unpopular speakers would likely not fare well.[143] We would risk leaving little space on platforms for a modern-day Galileo to share inconvenient truths or for protestors to engage in the digital equivalent of flag burning.

Under the second option, a company would be expected to justify speech restrictions based on the public interest determinations embodied in Article 19(3) without consideration of its economic incentives. The benefit of this option is that corporate speech codes would sync with international human rights law and seek to afford users the same (international law) protections as they have against governments, maintaining the scope of individual freedom of expression in the neo-medieval world. However, this option is unrealistic absent a substantial

---

companies to making public interest determinations when infringing on speech (i.e., the second option).

[143] Recalling the evolution of First Amendment speech protections is instructive in this regard. For well over a hundred years in the United States, interpretations of the appropriate scope of freedom of expression through democratically enacted laws resulted in bans on criticism of the government, slavery, and U.S. participation in wars. ANTHONY LEWIS, FREEDOM FOR THE THOUGHT WE HATE: A BIOGRAPHY OF THE FIRST AMENDMENT 23–38, 157–167 (2007). It was ultimately the courts (and not a decision by majority rule) that resulted in an interpretation of the First Amendment that protects unpopular, offensive, and minority views. *Id.*

societal shift regarding the special role of social media platforms as content moderators affecting discourse.

As Gillespie noted in his overview of content moderation, we "desperately need a thorough, public discussion about the social responsibility of platforms."[144] He reflects that this conversation unfortunately usually happens in the midst of particular controversies rather than with respect to the role of these platforms generally.[145] He states that platforms have tried to portray themselves as neutral "conduits, obscuring and disavowing the content moderation they do," when in reality they "invoke and amplify particular forms of discourse, and they moderate away others, all under the guise of being impartial conduits of open participation."[146]

A broad conversation on the role of social media platforms is essential to moving forward on the Special Rapporteur's recommendation to align corporate speech codes with human rights law. Such a conversation should explicitly force a reckoning about the basic trade-off that is at stake: with an ever-increasing amount of speech happening online in our neo-medieval world, should private platforms be able to develop substantive speech rules for any reason of their own choosing, or should individuals be able to enjoy the same rights to freedom of expression whether they are under the authority of a government or of a global social media platform? Perhaps such a conversation could trigger a societal shift from expecting companies ban speech as a measure of good corporate citizenship to building an expectation that good corporate citizenship means that platforms should respect international human rights standards when curating content.[147] If there were a growing consensus that platforms should respect the internationally recognized expression rights of users, then it would be possible that the economic incentives of companies would not undermine their ability to conduct public interest determinations. This could facilitate a path towards implementation of the second option: holding companies to the legitimacy prong of the tripartite test.

Assuming we could reach such a societal consensus, would companies then be well-positioned to make public interest determinations? From the GNI's experience, we know companies can assess whether governments are restricting speech based on a legitimate specified public interest reason or whether restrictions are invoked for illicit motives.[148] But this prong continues to pose potentially tricky questions when applied to corporations as judges of the public interest. If a government has made

---

[144] GILLESPIE, *supra* note 28, at 206.
[145] *Id.* at 206.
[146] *Id.* at 206–07.
[147] Such a broad conversation may also include consideration of whether some of the ills of cyberspace (hate speech, extremism, misinformation, etc.) can also be treated with societal interventions offline. Often to the extent that there is a conversation on the need for content moderation, the discourse (or news reporting) seems to stop at what platforms can do about the problem without considering what society at large can do about the issues.
[148] *See supra* notes 63–67 and accompanying text.

a public interest determination, should it be second-guessed by a company? For example, if a government does not believe hate speech on a platform is likely to lead to incitement to violence offline or otherwise does not risk the rights of others, could a social media company come to a different conclusion and properly justify a hate speech ban based on different public interest determinations? Should the level of deference by a company depend on how democratic the government actor is?[149] If a company assesses that a government is unable or unwilling to govern in the public interest, would it then be appropriate for the company to second-guess a government's public interest determination? Alternatively, should companies make their own public interest decisions regardless of determinations that have (or have not) been made by governments? Applying the third prong of Article 19(3)'s tripartite test raises a number of questions that would benefit from further conversations by interested stakeholders to assess the contours of what is feasible and to avoid corporations invoking public interest rationales as pretexts for revenue-driven decisions.

4. Observations on Applying the UNGPs to Corporate Speech Codes

This analysis concludes that the legality and necessity prongs of ICCPR Article 19(3)'s tripartite test can be adapted to the corporate context. Implementation of these two prongs with respect to speech codes would go a long way in helping to protect freedom of expression online. For example, in terms of the "legality" test, private speech codes could (and should) be modified to give concrete guidance rather than relying on vague prohibitions. Similarly, regarding the "necessity" test, companies should commit to engaging in the diligence required to select the least intrusive means of enforcing their speech codes. The third prong of the tripartite test, however, is the most difficult one to implement because, under the current state of affairs, expecting companies to disregard key economic motives in favor of the public interest seems unrealistic. Most

---

[149] Sometimes democratically-elected governments make decisions to limit speech that are not consistent with international human rights law. European approaches to limits on hate speech and extremism have recently been criticized by human rights groups and the UN Special Rapporteur on Freedom of Expression. *See, e.g.*, AMNESTY INT'L, DANGEROUSLY DISPROPORTIONATE: THE EVER-EXPANDING NATIONAL SECURITY STATE IN EUROPE 37–44 (2017) (criticizing several European countries for laws with vague prohibitions, such as the "glorifying" or "promoting" of terrorism); David Kaye (Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression), *Rep. of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 25, U.N. Doc. A/71/373 (Sept. 6, 2016) (criticizing European human rights law for failing to "define hate speech adequately"). *See also* Kristen Eichensehr, *Digital Switzerlands* 167 U. PA. L. REV (forthcoming 2019) (manuscript at 41) (https://poseidon01.ssrn.com/delivery.php?ID=963020092123086109009087126069016073033078040100220060940751260021021130110241250070060580390441111130281250000861220690031111230820690480920961200710901100940310350930151221061001251100650091221221091130801231260300890300690701241090010070841201111&EXT=pdf) ("[C]ompanies will 'fold'—complying with rather than challenging, government requests—when they perceive governments and users to be aligned.").

likely they would invoke public interest determinations as pretexts for revenue-driven decisions. That said, if there is a societal shift to expecting platforms to respect international freedom of expression protections online, it may be more feasible for companies to make public interest determinations. But, questions remain that would benefit from additional multi-stakeholder deliberations about implementation of this prong.

III. REFLECTIONS ON POTENTIAL CRITICISMS AND BENEFITS OF HUMAN RIGHTS LAW AS THE DEFAULT RULE FOR CORPORATE SPEECH CODES

Having examined the type of revisions to corporate speech codes that would be triggered if companies align them with ICCPR Article 19, this analysis next turns to assessing potential criticisms and benefits of the Special Rapporteur's proposed approach. A range of potential criticisms and concerns are examined, including whether international human rights law provides companies with adequate guidance in regulating speech, whether U.S. companies should promote First Amendment standards instead, and potential adverse impacts companies could have on the international human rights regime. The potential benefits that are considered include improved free speech protections for individuals in a neo-medieval world, a principled basis for companies to regulate speech worldwide, and the fact that the framework to implement this approach is already in place. This Article concludes that the benefits of progressing towards alignment of corporate speech codes with international human rights law outweigh the potential downsides.

A. Criticisms

The Special Rapporteur's recommended approach has already been criticized. For example, one commentator questioned the Special Rapporteur's proposed approach because "[i]t is something of a misnomer to speak of international human rights law as if it is a single, self-contained and cohesive body of rules. Instead, these laws are found in a variety of international and *regional* treaties that are subject to differing interpretations by states that are parties to the conventions as well as international tribunals applying the laws."[150] Such a concern inappropriately conflates international human rights law with separate bodies of law embodied in regional human rights instruments.[151] The

---

[150] Evelyn Douek, *U.N. Special Rapporteur's Latest Report on Online Content Regulation Calls for 'Human Rights by Default*,' LAWFARE: BLOG (June 6, 2018, 8:00 AM), https://www.lawfareblog.com/un-special-rapporteurs-latest-report-online-content-regulation-calls-human-rights-default (emphasis added).

[151] *See supra* notes 90–97 and accompanying text for a discussion of differences between the international human rights regime and regional regimes with regard to freedom of expression. When there are areas of overlap between these systems, international and regional human rights mechanisms will occasionally make joint statements on high profile topics, but that does not mean the international and regional systems are the same or congruous in every regard. *See, e.g.*, UNITED NATIONS (UN) SPECIAL RAPPORTEUR ON FREEDOM OF OPINION & EXPRESSION, ORG. FOR SEC. & CO-OPERATION IN EUR. (OSCE) REPRESENTATIVE ON FREEDOM

differences between international and regional human rights law do not make international human rights law's protection for freedom of expression internally inconsistent; it just means international and regional human rights law are separate bodies of law that do not always align.[152] The call of the UNGPs (and the Special Rapporteur) is for companies to respect international, not regional, human rights law. Under international human rights law, the key protection for speech comes from Article 19(3)'s tripartite test, which applies to all speech restrictions.[153]

Another potential concern is whether the international human rights law regime on freedom of expression provides sufficient guidance to companies in applying speech restrictions.[154] Of course, this international human rights regime on speech is not a detailed tax code setting forth a comprehensive listing of unprotected terms or phrases for the entire world. The inherent nature of speech adjudications requires

---

OF THE MEDIA, & ORG. OF AM. STATES (OAS) SPECIAL RAPPORTEUR ON FREEDOM OF EXPRESSION, & AFRICAN COMM'N ON HUMAN & PEOPLES' RIGHTS (ACHPR) SPECIAL RAPPORTEUR ON FREEDOM OF EXPRESSION & ACCESS TO INFO., JOINT DECLARATION ON FREEDOM OF EXPRESSION AND 'FAKE NEWS', DISINFORMATION AND PROPAGANDA (Mar. 3, 2017), https://www.osce.org/fom/302796?download=true (commemorating that UN and regional free expression experts endorse several approaches to combatting false news, such as affirming ICCPR Article 19's tripartite test for any restrictions, condemning broad bans on "fake news" as unduly vague, and calling for an end to the criminalization of defamation); *Joint Declaration on Freedom of Expression and Countering Violent Extremism*, OFFICE OF THE HIGH COMM'R FOR HUMAN RIGHTS, https://www.ohchr.org/En/NewsEvents/Pages/DisplayNews.aspx?NewsID=19915&LangID=E (last visited Aug. 10, 2018) (agreeing on a number of measures to combat terrorism that inappropriately restrict speech, including vague bans on "glorification of terrorism" and "apology for terrorism").

[152] It should be noted that nations cannot invoke interpretations of regional human rights mechanisms to get out of their international obligations. For example, Germany cannot invoke the case law of the European Court of Human Rights approving of bans on Holocaust denial under the European Convention on Human Rights to justify its Holocaust denial bans under the ICCPR. *See supra* notes 92–93 and accompanying text, for the differences between the two systems with regard to denials of historic atrocities.

[153] *See supra* note 47 and accompanying text (explaining that even mandatory speech bans in international treaties are subject to ICCPR's Article 19(3)'s tripartite test).

[154] *See, e.g.*, Citron, *supra* note 140, at 1063 (dismissing international human rights law as a source of guidance for tech companies in defining hate speech and terrorist-related speech because "human rights law contains exceptionally flexible standards" and recommending companies look to European approaches). It should be noted that the UN Special Rapporteur on freedom of expression has criticized the vagueness of European human rights law with respect to hate speech. Kaye, *supra* note 149. *See also* David Kaye, *How Europe's New Internet Laws Threaten Freedom of Expression: Recent Regulations Risk Censoring Legitimate Content*, FOREIGN AFFAIRS (Dec. 18, 2017), https://www.foreignaffairs.com/articles/europe/2017-12-18/how-europes-new-internet-laws-threaten-freedom-expression (describing a wave of European regulations, including with respect to terrorist material and hate speech, that "risk interfering with" international freedom of expression standards).

investigations into context and related judgement calls. ICCPR Article 19(3)'s tripartite test does provide a rigorous, principled, and coherent standard that allows for such judgment calls in considering restrictions on speech in each country. Despite the fact that there is not a UN court dedicated to issuing legally binding decisions on proper interpretations of the ICCPR, the UN's human rights machinery has provided a substantial amount of guidance and recommendations in interpreting this article. For example, after requesting the views of civil society and State Parties, in 2011 the UN Human Rights Committee issued lengthy guidance about Article 19, ranging from issues of defamation to restrictions based on national security to access to information.[155] The Special Rapporteur position has existed since 1993 and has issued numerous reports and guidance on a variety of issues arising under Article 19.[156] Any criticism that Article 19 does not provide sufficient guidance seems to overlook the body of recommendations by UN independent experts on this topic. If companies were to accept the call to regulate speech in line with ICCPR Article 19(3)'s tripartite test, they would be using an internationally accepted and principled standard that gives space for consideration of context in making the judgment calls inherent in speech adjudications. They would also find the recommendations of the Human Rights Committee and the Special Rapporteur valuable in implementing such an approach.

Another criticism could stem from the fact that some may prefer that American companies curate speech on their platforms in accordance with the First Amendment, which provides one of the most robust protections for speech in the world.[157] The founders, leading officers, and legal teams of many prominent U.S. social media companies seem to have been heavily influenced by First Amendment principles, particularly at the outset of the companies' operations.[158] However, the speech codes for

---

[155] GC 34, *supra* note 39, ¶¶ 1–52.

[156] *Freedom of Opinion and Expression – Annual Reports*, OFFICE OF THE HIGH COMM'R FOR HUMAN RIGHTS, https://www.ohchr.org/EN/Issues/FreedomOpinion/Pages/Annual.aspx (last visited Aug. 10, 2018) (analyzing issues ranging from encryption and anonymity to regulation of online content to the treatment of whistleblowers).

[157] *See, e.g.*, David French, *A Better Way to Ban Alex Jones*, N.Y. TIMES (Aug. 7, 2018), https://www.nytimes.com/2018/08/07/opinion/alex-jones-infowars-facebook.html?action=click&pgtype=Homepage&clickSource=story-heading&module=opinion-c-col-right-region&region=opinion-c-col-right-region&WT.nav=opinion-c-col-right-region (criticizing tech companies for abandoning First Amendment principles and using subjective standards to ban Alex Jones from their platforms). *See also* Noah Feldman, *Free Speech Isn't Facebook's Job*, BLOOMBERG: BLOOMBERGOPINION (June 1, 2016, 11:08 AM), https://www.bloomberg.com/view/articles/2016-06-01/it-s-not-facebook-s-job-to-guarantee-free-speech (expressing outrage initially at tech companies for selling out on First Amendment principles but concluding society cannot expect companies to respect freedom of expression).

[158] Klonick, *supra* note 22, at 1618–25 (finding American lawyers trained in First Amendment jurisprudence to have heavily influenced the initial speech codes and approaches of leading platforms); Ammori, *supra* note 22, at 2283.

such companies have steadily moved away from U.S. constitutional free speech protections.[159] Simply put, we are no longer in a situation in which prominent companies would be abandoning speech codes that reflect First Amendment approaches in favor of international approaches.[160]

Moreover, a rigorous and good faith interpretation of ICCPR Article 19's tripartite test of legality, necessity, and legitimacy would bring company speech codes much closer to First Amendment standards than what is currently happening with the curation of speech on platforms. As previously noted, many company speech codes are vague.[161] Corporate implementation of ICCPR speech protections means companies would need to revise their codes to give users appropriate notice of the parameters of unacceptable speech. The speech codes would also need to be adjusted so as to not discriminate against any group.[162] Often the combination of having to overcome vague terminology and avoid discrimination against any group makes it difficult to craft broad speech bans.[163] In addition, corporations would need to commit to selecting enforcement options that reflect the least intrusion on speech interests to be consistent with the "necessity" prong of the tripartite test.[164] Thus, the proper application of at least the first two prongs of Article 19(3) coupled with other ICCPR protections, such as the ban on discrimination, would serve as a principled check on corporate speech bans if applied in good faith.[165]

Another potential critique is that the call to align company speech codes with international human rights law is based on a framework that is not legally binding, i.e., the UNGPs. This means that grounding speech codes in international human rights law would be a voluntary action taken by companies to live up to the international community's expectations. Can we entrust the future of freedom of expression online to the mere hope that companies will voluntarily implement the UNGPs? Perhaps we have

---

[159] Ammori, *supra* note 22, at 2274–84 (describing how speech codes of U.S. platforms depart from First Amendment principles but are influenced by the First Amendment); Citron, *supra* note 140 (describing how the European Union pressured American companies to change their approaches to hate speech and terrorist material).

[160] If prominent social media companies had displayed a commitment to grounding their speech codes in the First Amendment despite pressure from advertisers and the public, this analysis of the concerns and benefits of aligning corporate speech codes with international human rights law would be significantly different.

[161] *See supra* note 110 and accompanying text.

[162] *See supra* note 42 and accompanying text.

[163] *See* ERWIN CHEMERINSKY & HOWARD GILLMAN, FREE SPEECH ON CAMPUS 93–94 (Yale University Press 2017) (assessing that U.S. courts often do not uphold bans on fighting words because such prohibitions are either vague or show favoritism towards particular groups).

[164] *See supra* text accompanying notes 112–135.

[165] It may also be challenging for a company to justify its speech restriction under Article 19's three prong test as "necessary," for example, to maintain public order if a government (e.g., the United States) has not assessed there is a public order problem that justifies speech restrictions. *See supra* notes 148–149 and accompanying text.

to focus on voluntary corporate initiatives for at least two reasons. First, intervention by the U.S. government is highly unlikely given that First Amendment protections for corporate speech provide the United States with little room for legally binding solutions. Second, an international negotiation to regulate speech on platforms, including content moderation, is undesirable because it would no doubt be dominated by powerful countries with weak records on freedom of expression that would seek to roll back international speech protections. To begin with, global trends show governments have become more restrictive with respect to online speech, which means such countries would seek to commemorate their problematic approaches to online speech in any new international instrument.[166] Such trends coupled with the recent withdrawal by the United States, a traditional global leader in promoting robust free expression norms in multilateral fora, from the UN Human Rights Council (which would likely negotiate any such agreement) means the prospects for any new international treaty protecting speech as robustly as ICCPR Article 19(3) does are bleak.[167]

Moreover, the trajectory of the business and human rights movement has been positive (though not swift) with companies increasingly undertaking measures to align their operations with international human rights standards on a voluntary basis. A 2016 study found that of 275 General Counsels and senior lawyers surveyed, forty-six percent of businesses have human rights policies.[168] For companies making more than ten billion dollars in revenue, eighty-four percent have human rights policies.[169] As noted previously, within the information and communication technology sector, companies in the GNI such as Google, Facebook, Microsoft, and Oath (the successor to Yahoo! and America Online) have opted to respond to worldwide governmental requests to restrict speech in ways that seek to avoid infringements on international human rights.[170] Oath has a specialized business and human rights unit focused on expression and privacy.[171] Many of these companies (and others) voluntarily issue transparency reports regarding requests from governments that infringe on expression and privacy and the corporate

---

[166] *See, e.g.*, SANJA KELLY ET AL., FREEDOM HOUSE, FREEDOM ON THE NET 2017: MANIPULATING SOCIAL MEDIA TO UNDERMINE DEMOCRACY (Nov. 2017), *available at* https://freedomhouse.org/sites/default/files/FOTN_2017_Final.pdf.

[167] *See* Susan Hannah Allen & Martin S. Edwards, *The U.S. Withdrew from the U.N. Human Rights Council. That's Not How the Council Was Supposed to Work,* WASH. POST (June 26, 2018), https://www.washingtonpost.com/news/monkey-cage/wp/2018/06/26/the-u-s-withdrew-from-the-u-n-human-rights-council-thats-not-how-the-council-was-supposed-to-work/?utm_term=.83478fe1cf27 (discussing U.S. withdrawal from the Council and its implications).

[168] James Wood, *The New Risk Front for GCs—Nearly Half of Contracts Have Human Rights Clauses, LB Research Finds*, LEGAL BUS.: BLOG (Sept. 8, 2016, 8:46 AM), https://www.legalbusiness.co.uk/blogs/the-new-risk-front-for-gcs-nearly-half-of-contracts-have-human-rights-clauses-lb-research-finds/.

[169] *Id.*

[170] *See supra* text accompanying notes 63–67.

[171] *Business & Human Rights at Oath*, OATH, https://www.oath.com/our-story/business-and-human-rights/ (last visited Aug. 10, 2018).

actions taken in response.[172] The fact that social media companies are searching for a legitimate basis to regulate speech around the world[173] should also incentivize corporate action towards this already established global standard embodied in ICCPR Article 19.

Yet another potential criticism with regard to encouraging companies to align their speech codes with international human rights law is the feasibility of such an endeavor given the scale at which the companies operate. No government has had to implement its human rights obligations at the scale at which these global platforms operate. This Article does not seek to dismiss how challenging it is for companies to administer their speech codes on a global basis.[174] That said, companies have already decided to have complex speech rules and are already applying them globally. They appear to realize that they need more staff and better procedures to implement their existing codes.[175] Others have already argued eloquently for better procedural safeguards and transparency measures in corporate content moderation.[176] The shift towards grounding the speech codes in international human rights law merely seeks to anchor the existing global speech curation process to speech codes that are consistent with international standards for restricting speech, rather than to speech codes that are "homegrown" approaches to restricting speech.

Though this section is not exhaustive in terms of potential criticisms, perhaps two more bear mentioning. With U.S. abandonment of

---

[172] Several prominent social media companies issue transparency reports concerning governmental requests to remove speech from their platforms, but YouTube became "the first major social media platform to put out a report on the number of posts it removes under its own content policy" in April 2018. Liz Woolery, *Companies Finally Shine a Light into Content Moderation Practices*, CTR. FOR DEMOCRACY & TECH.: BLOG (Apr. 25, 2018), https://cdt.org/blog/companies-finally-shine-a-light-into-content-moderation-practices/.

[173] *See infra* notes 186–189 and accompanying text.

[174] For a discussion of the complexity of global content moderation processes, see Klonick, *supra* note 22, at 1631–48 and GILLESPIE, *supra* note 28, at 111–72.

[175] *See supra* note 27 and accompanying text; *How Social-Media Platforms Dispense Justice*, THE ECONOMIST (Sept. 6, 2018), https://www.economist.com/business/2018/09/08/how-social-media-platforms-dispense-justice?fsrc=scn/tw/te/bl/ed/howsocialmediaplatformsdispensejusticethedeciders (reporting that by the end of 2018 "Facebook will have doubled the number of employees and contractors dedicated to the 'safety and security' of the site, to 20,000, including 10,000 content reviewers. YouTube will have 10,000 people working on content moderation in some form."). While some companies have been increasing the number of content moderators, decisions continue to be made with extraordinary rapidity, which is highly problematic given the time needed for human judgment in speech adjudication. *See* GILLESPIE, *supra* note 28, at 121 ("Fast here can mean mere seconds per complaint – approve, reject, approve – and moderators are often evaluated on their speed as well as their accuracy, meaning there is reward and pressure to keep up…. Each complaint is thus getting just a sliver of human attention, under great pressure ….").

[176] *See supra* note 35.

its seat at the Human Rights Council, there are serious risks of backsliding at the United Nations on freedom of expression protections, as those opposed to this right may become more active in future resolutions involving freedom of expression.[177] Although Council resolutions are not legally binding, they can reflect at times an important political consensus on speech issues that can impact the rest of the UN's human rights machinery.[178] Not participating at the Human Rights Council will also prevent the United States from having a persuasive voice with respect to the selection of the next Special Rapporteur on freedom of expression, which could negatively impact future developments on this topic. In addition, the United States is generally able to have one of its citizens elected to the UN Human Rights Committee, and that independent expert traditionally brings important U.S. perspectives to the Committee's recommended interpretations of the ICCPR. Recently, the U.S. candidate was not elected, which means the Human Rights Committee will not have an independent expert who can bring American experiences and perspectives to its work.[179]

This combination of factors could result in a roll-back of freedom of expression protections at the international level. Though the Human Rights Committee and Special Rapporteur's existing guidance should help to temper such backsliding, regressive recommendations about the scope of this right could happen.[180] If it does, that would be a significant drawback to linking corporate speech codes to international human rights law. However, if American companies (with First Amendment roots and inclinations) become active and effective stakeholders in trying to promote broad protections for speech under international human rights law, their influence could potentially serve as a check on such regression in the absence of U.S. leadership at the Human Rights Council.[181]

---

[177] Peter Micek, *Saving the U.N. "Internet Resolution" from Sharks Circling in Geneva*, ACCESS NOW (July 10, 2018, 7:27 PM), https://www.accessnow.org /saving-the-u-n-internet-resolution-from-sharks-circling-in-geneva/ ("Normally, the U.S. would have been a key member state working on this [Internet] resolution. In previous years, the U.S. has been part of the 'core group' of authors, and has co-sponsored the text. But this year the U.S. was absent, having withdrawn from the Human Rights Council just as the negotiations for this resolution began. . . . Protecting human rights is difficult and messy work . . . and leaves people who cannot protect themselves even more vulnerable. If the absence of the U.S. emboldened states seeking more control over the internet, the lesson here is clear: those truly committed to human rights must engage more deeply.").

[178] For example, the Special Rapporteur on freedom of opinion and expression often cites to these resolutions. *See*, *e.g.*, Kaye, *supra* note 149, ¶¶ 22, 33, 41, 45.

[179] Barbara Crossette, *The UN Eyes a World with Less US*, THE NATION (July 30, 2018), https://www.thenation.com/article/un-eyes-world-less-us/.

[180] One way forward could be to link corporate speech codes to existing guidance from the UN's human rights machinery to avoid negative impacts if regressive recommendations emerge in the future.

[181] For an interesting discussion of the potential for the rise of U.S. technology companies to serve as a check on governments, see Eichensehr, *supra* note 149, at 49 ("[H]aving two powerful regulators, rather than only one, can benefit individuals' freedom, liberty, and security because sometimes it takes a powerful

Conversely, there could be concerns about the potential impact that corporate implementation of international human rights norms could have on the international human rights regime itself if companies do not interpret ICCPR Article 19 in good faith. Given that an international human rights court solely dedicated to adjudicating ICCPR rights does not exist, the UN machinery's recommended interpretations of UN standards have come primarily from the Human Rights Committee, the Special Rapporteur, and (occasionally) certain high-profile, non-binding consensus resolutions adopted by UN member states.[182] If companies begin applying Article 19(3) in their content moderation operations and take up the Special Rapporteur's call to produce "case law," there could be an active fountain of new "jurisprudence" involving the ICCPR's speech protections, which could influence the direction of international freedom of expression rights. It is thus even more important that companies apply the international standards in good faith rather than in revenue-driven ways that could undermine the robustness of the standards with respect to state actors. This seems to be a risk that is worth taking in order to afford ICCPR protections for users' speech when they are under the authority of platforms.[183] The alternative is to leave individuals under speech codes that are untethered to any traditional sources of restraint in regulation, i.e., the First Amendment or international human rights law.

B. Benefits

A number of significant benefits exist to grounding the substantive restrictions of company speech codes in international human rights law. First and foremost, anchoring corporate speech codes to ICCPR Article 19 represents the best chance of protecting the freedom of expression interests for users throughout the world. As previously discussed, companies currently set substantive speech codes as they see fit. While they may have started out heavily influenced by the First Amendment, their codes have steadily moved away from that standard due to revenue concerns, public pressure, and governmental pressure to re-interpret their terms of service in a more restrictive way.[184] As the U.S. government is unlikely to regulate the speech codes of private companies given constitutional protections for corporate speech rights and international regulation of such

---

regulator to challenge and check another powerful regulator."). Admittedly, the role of strengthening the human rights regime seems to go beyond what is called for in the UNGPs, but it would be consistent with the companies' mission statements to promote the free flow of information and self-interest to promote broad expression rights online.

[182] *See, e.g.*, Human Rights Council Res. 20/8, U.N. Doc. A/HRC/RES/20/8, ¶ 1 (July 16, 2012) (affirming for the first time that individuals have the same rights online as they have offline); Council Res. 16/18, *supra* note 115 and accompanying text.

[183] Some of the suggestions proposed by the Special Rapporteur's report—involving improved transparency, procedures, and oversight for content moderation—may be helpful in assessing whether companies are respecting the standards in ICCPR Article 19's tripartite test. *See SR Report*, *supra* note 79.

[184] *See supra* notes 139–140 and accompanying text.

codes is undesirable,[185] seeking to have companies align their speech codes with international human rights law remains the best avenue for protecting individuals' speech rights in our neo-medieval world. The alternative would be for each company to develop its own code based on its own views of speech and revenue needs, which is not a stable foundation for the long-term protection of speech. The fact that Twitter and Facebook recently expressed an openness to considering international human rights law in their speech codes also gives momentum to this path.[186]

Aligning speech codes with the ICCPR also has a number of benefits for companies. Major platforms appear to be seeking a principled basis for regulating speech in every country of the world in order to give legitimacy to their global content moderation. For example, Facebook CEO Mark Zuckerberg stated:

> With a community of more than 2 billion people all around the world, in every different country, where there are wildly different social and cultural norms, it's just not clear to me that us sitting in an office here in California are best placed to always determine what the policies should be for people all around the world. And I've been working on and thinking through: How can you set up a more democratic or community-oriented process that reflects the values of people around the world? That's one of the things that I really think we need to get right. Because I'm just not sure that the current state is a great one.[187]

Similarly, a lawyer working for Google in 2006 was tasked with figuring out how to respond to the Thai government's demand to remove offensive YouTube videos of the king.[188] After meeting with Thai people and observing how shaken ordinary citizens were by these insults to their king, she felt "Who am I, a U.S. attorney sitting in California to tell them: 'No, we're not taking that down.'"[189] She and her team removed the videos from view within Thailand.[190]

Companies need not recreate the wheel in developing speech norms that have worldwide legitimacy if they base their content moderation policies on international human rights standards. Since 1966, there has been an international treaty (the ICCPR) that protects freedom of expression with an international machinery for monitoring its implementation. Aligning company speech codes with existing international human rights law would give companies a legitimate, principled, and international basis upon which to make decisions that affect freedom of expression throughout the world. For example, instead

---

[185] *See supra* notes 166–167 and accompanying text.
[186] *See supra* notes 33–34 and accompanying text.
[187] Ezra Klein, *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next*, VOX (Apr. 2, 2018, 6:00 AM), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.
[188] Klonick, *supra* note 22, at 1623.
[189] *Id.*
[190] *Id.*

of struggling for a way to justify his decision to permit Holocaust denial posts on his platform, Mr. Zuckerberg could have cited to the UN Human Rights Committee's interpretation of ICCPR Article 19.[191] Similarly, in the case of the YouTube videos mocking royalty in Thailand, a corporate decision grounded in this Committee's recommendations[192] might have appeared more principled to Thai citizens than what they were left with—the views of lawyers in Silicon Valley.

Companies would also benefit from linking their speech codes to international human rights law because countries often pressure companies (1) to interpret their own terms of service in a restrictive manner, or (2) to remove content from their platforms that conflicts with local law but would otherwise be protected by international human rights law. Grounding corporate speech codes in Article 19 of the ICCPR can help companies better resist such measures under either situation. For example, if Europe pressures tech companies to interpret their hate speech codes loosely or to remove illegal hate speech under unrealistic time frames, companies could push back by saying their codes are aligned with international human rights law and thus cannot be interpreted or implemented in such a fashion. Similarly, it places companies in an untenable spot to say to governments: "We will not remove speech critical of the government because you need to respect users' international freedom of expression rights, but we can certainly remove that content if we feel like it." Companies will be on firmer ground to resist illicit governmental demands and laws if they treat user speech as protected under the same rubric for corporate speech codes and governmental regulation.

Aligning corporate speech codes with international human rights law protections, which is what the UNGPs call for, has the added benefit of providing a way forward that does not require international negotiations

---

[191] The UN Human Rights Committee has taken the position that laws that restrict opinion about historical facts are an unacceptable infringement on speech. GC 34, *supra* note 39, ¶ 49 ("Laws that penalize the expression of opinions about historical facts are incompatible with the obligations that the Covenant imposes on States parties in relation to the respect for freedom of opinion and expression. The Covenant does not permit general prohibition of expressions of an erroneous opinion or an incorrect interpretation of past events.").

[192] The UN Human Rights Committee has criticized laws that protect royalty or heads of state from criticism. GC 34, *supra* note 39, ¶ 38 ("[T]he Committee has observed that in circumstances of public debate concerning public figures in the political domain and public institutions, the value placed by the Covenant upon uninhibited expression is particularly high. Thus, the mere fact that forms of expression are considered to be insulting to a public figure is not sufficient to justify the imposition of penalties, albeit public figures may also benefit from the provisions of the Covenant. Moreover, 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. Accordingly, the Committee expresses concern regarding laws on such matters as, lese majesty, *desacato*, disrespect for authority, disrespect for flags and symbols, defamation of the head of state and the protection of the honour of public officials . . . .").

about how to tackle the rise of private sector regulation of speech. The UNGPs already exist, were endorsed by consensus at the Human Rights Council, and reflect the international community's expectations that companies will respect human rights in all their operations, including content moderation. Similarly, the ICCPR already exists and rigorous application of Article 19(3) provides a strong check against inappropriate restrictions on speech. Engaging in international negotiations to develop a way forward with respect to transnational private sector content moderation creates an unacceptable risk of regression in free expression rights for a variety of reasons previously discussed.[193] In sum, there is significant value to using an international regime that already exists, has global approval, and that, if applied properly, would result in corporations respecting the international freedom of expression rights of users throughout the world.

C. Observations on Criticisms and Benefits

Overall, while some potential pitfalls exist to anchoring corporate speech codes to international human rights law, the benefits seem to outweigh such downsides. In particular, this approach appears to be the most viable route to promote corporate respect for individuals' freedom of expression rights in a neo-medieval world. In addition, this approach would likely increase companies' legitimacy in content moderation while also help companies resist demands from governments to restrict speech in ways at odds with international human rights law. Given that we appear to be in a unique norm-setting moment in the thinking of platforms with regard to the substantive content of their speech codes, this approach provides the best available way forward for users and companies.

IV. CONCLUSION

Recent events, such as Facebook's removal of a post that contained an "offensive" part of the Declaration of Independence and its subsequent decision to permit Holocaust denial posts on its platform as well as the decision of many tech giants to de-platform a conspiracy theorist, have highlighted the enormous power of corporate actors over freedom of expression. Though much of the commentary to date has focused on the need for platforms to increase transparency and add procedural safeguards for users when moderating content, the summer of 2018 seemed to mark a norm-setting opportunity for the substantive content of corporate speech codes. In June 2018, the UN Special Rapporteur on expression called on companies to align their speech codes

---

[193] *See supra* notes 166–167, 177. Also, on the domestic level, the constant debates about whether to treat platforms as utilities, publishers, or something else are also not likely to reach a resolution that would resolve the substantive issues of corporate speech codes in time to affect the existing norm setting moment. For a discussion of the legal issues involved, see Klonick, *supra* note 22, at 1660–63.

with international human rights law. After the controversy surrounding whether to de-platform Alex Jones, two social media giants seemed open to considering international human rights law as a basis for their speech decisions.

This Article set out to analyze what it would mean in practice for such companies to align their speech codes with international human rights law: specifically, whether it is *feasible* for companies to do so and whether such an outcome is *desirable*. This Article began with an overview of applicable international human rights law standards. The ICCPR, the most relevant treaty on freedom of expression, provides broad protections for speech across borders, but permits restrictions on speech if every prong of Article 19(3)'s tripartite test is met. The prongs are as follows: any restriction (1) must not be vague, and (2) must constitute the "least intrusive means" to (3) achieve a legitimate public interest. Under international human rights law, any restriction on speech must meet this tripartite test, even those restrictions imposed under treaty provisions that mandate barring incitement to violence and other harms. While Article 19 is not directly applicable to companies (as they are private actors and not states), the 2011 UNGPs reflect the international community's expectation that companies will arrange their business operations, including their terms of service, to respect international human rights.

In order to assess the feasibility of aligning corporate speech codes with the ICCPR, this Article focused on a concrete example: Twitter's hate speech rules. With respect to the first prong of Article 19(3)'s tripartite test, the analysis found that Twitter's rules were vague and would need to be revised to give users more notice of what is not allowed. In analyzing the second prong of the tripartite test in the context of a corporate actor, this Article argued that a company should (1) evaluate the means at its disposal to achieve a legitimate aim without infringing on speech; (2) select enforcement options for speech code violations that least intrude on speech interests; and (3) periodically assess whether the selected measure helps to achieve the legitimate aim. This Article noted that Twitter should commit publicly to using the least intrusive enforcement actions to deal with speech code violations. With respect to the third prong—regulating speech for the public interest—this Article observed that companies generally regulate speech based on revenue-related motivations, which could make this prong challenging to implement in good faith. This Article recommended a broad societal conversation to clarify the role of platforms in the protection of speech, which would help facilitate public interest determinations. Overall, aligning corporate speech codes with much of ICCPR Article 19's tripartite test is feasible, but further discussion is needed with respect to the role of companies in making public interest determinations.

In considering the desirability of having companies align their speech codes with international human rights law, this Article also considered a variety of potential criticisms of such an approach. It concluded that arguments of the alleged incoherency of international human rights law often inappropriately conflate the international human rights regime with regional regimes rather than reflecting an analysis of

the ICCPR. This Article determined that criticisms of human rights law not providing sufficient guidance to companies actually do not fully grasp how ICCPR Article 19(3)'s tripartite test works, the recommended interpretations of the UN's human rights machinery, or the fact that speech adjudications inherently involve judgment calls that consider contextual factors. To the extent commentators would prefer American companies to enforce First Amendment principles in content moderation, this Article noted that the speech codes of prominent platforms no longer reflect such principles, and a proper application of Article 19(3)'s tripartite test would likely protect more speech than is currently the case. While recognizing that the UNGPs reflect the international community's expectations but do not constitute a legally binding framework, this Article noted the trajectory of the business and human rights movement has been positive, and the corporate interests in adjudicating speech based on universally accepted standards could incentivize voluntary adoption of the Special Rapporteur's recommended approach. The analysis also noted that U.S. regulation is unlikely and international regulation is undesirable, as it would likely result in a substantial diminution of international free speech protections. But the analysis expressed concerns about the implications for future developments at the international level on freedom of expression given U.S. withdrawal from the UN Human Rights Council.

This Article ultimately concluded that the advantages of aligning corporate speech codes with international human rights law outweigh the potential disadvantages. This approach seems to be the most feasible way to protect against infringements on users' freedom of expression rights by private actors. The approach should also be appealing to companies who seem to be grappling to find a principled basis upon which to regulate speech worldwide, as well as a principled basis on which to resist governmental demands that violate freedom of expression. The fact that international human rights law and the UNGPs reflect an international consensus is also a valuable aspect of this approach, as it avoids lengthy and potentially problematic international negotiations involving corporate speech codes. In our neo-medieval world, the most viable way to protect individuals' freedoms of expression rights is to seek to have governments implement their international human rights obligations regarding speech and encourage companies to align their codes with these standards.

Since Eleanor Roosevelt led the U.S. delegation in negotiating the ICCPR, U.S. diplomats have consistently fought in UN fora to maintain broad speech protections under international law. With the rise of powerful corporate actors engaging in a parallel governance exercise over speech alongside governments throughout the world, it is important for these companies to recognize that in many ways they have been handed the baton of respecting and promoting international freedom of expression protections. American platforms may not have asked to be in this position, but this is an important norm-setting moment in which tech giants could greatly and positively influence the future of freedom of expression online. They should acknowledge their roles as powerful co-regulators of speech and hold themselves to the same protections for freedom of expression that apply to state actors. We should be encouraging companies to respect international human rights in our brave neo-medieval world or face a

future in which their speech codes are untethered to any speech-protective norms.