Journal of Technology Law & Policy

Volume XX – 2019-2020

ISSN 2164-800X (online)

DOI 10.5195/tlp.2020.236

http://tlp.law.pitt.edu

# Squaring the Circle Between Freedom of Expression and Platform Law

Michael Karanicolas

Abstract

Among the greatest emerging challenges to global efforts to promote and protect human rights is the role of private sector entities in their actualization, since international human rights rules were designed to apply primarily, and in many cases solely, to the actions of governments. This paradigm is particularly evident in the expressive space, where private sector platforms play an enormously influential role in determining the boundaries of acceptable speech online, with none of the traditional guardrails governing how and when speech should be restricted. Many governments now view platform-imposed rules as a neat way of sidestepping legal limits on their own exercise of power, pressuring private sector entities to crack down on content which they would be constitutionally precluded from targeting directly. For their part, the platforms have grown increasingly uncomfortable with the level of responsibility they now wield, and in recent years have sought to modernize and improve their moderation frameworks in line with the growing global pressure they face. At the heart of these discussions are debates around how traditional human rights concepts like freedom of expression might be adapted to the context of "platform law." This Article presents a preliminary framework for applying foundational freedom of expression standards to the context of private sector platforms, and models how the three-part test, which lies at the core of understandings of freedom of expression as a human right, could be applied to platforms' moderation functions.



This site is published by the University Library System of the University of Pittsburgh as part of its D-Scribe Digital Publishing Program and is cosponsored by the University of Pittsburgh Press.

Table of Contents





Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

# Squaring the Circle Between Freedom of Expression and Platform Law

Michael Karanicolas*

## INTRODUCTION

Privatization can be a controversial practice. To its proponents, it is an engine of efficiency, introducing a competitive atmosphere to stodgy and self-perpetuating bureaucracies. However, there are externalities which come into play when governments abrogate direct responsibility over an area of administration. A private prison may be run at less cost to the taxpayer, but will it respect the rights of inmates and devote sufficient resources to their rehabilitation? Privatizing a water company could turn it profitable, but this might come at the cost of an increase in contaminants, or a refusal to properly service rural areas. Despite the common trope that "government should be run like a business,"[1] there is an important distinction between the core functions of these two types of entities. A private company's core purpose is to maximize profit for its shareholders. A government's core purpose is to promote and protect the rights of its people.[2]

Determining how and where to regulate speech is among the most important, and most delicate, tasks a government may undertake.[3] It requires a careful balancing between removing harmful content while providing space for controversial and challenging ideas to be aired, and deterring dangerous speech while minimizing a

* Resident Fellow at Yale Law School, leading the Wikimedia Initiative on Intermediaries and Information. Sincere thanks to Anna Su, Matthew Marinett, Jack Enman-Beech, Przemysław Pałka, Thomas Kadri, Chinmayi Arun, Maren Woebbeking, Anat Lior, and Evelyn Douek, all of whom offered very helpful feedback in revising this Article.

[1] Philip Bump, *Trump's Idea to Run the Government Like a Business is an Old One in American Politics*, WASH. POST (Mar. 27, 2017), https://www.washingtonpost.com/news/politics/wp/2017/03/27/trumps-idea-to-run-the-government-like-a-business-is-an-old-one-in-american-politics/?noredirect=on&utm_term=.6f0808fab57d.

[2] MORTON E. WINSTON, THE PHILOSOPHY OF HUMAN RIGHTS 9 (Kenneth M. King et al. eds., Belmont, CA: Wadsworth 1989).

[3] *See, e.g.*, Irwin Toy Ltd. v. Quebec (A.G.), [1989] 1 S.C.R. 927, 976; *see also* Ford v. Quebec, [1988] 2 S.C.R. 712, 765–66.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

broader chilling effect that impacts legitimate areas of debate.[4] The challenges in regulating expression are among the most vibrant and hotly debated areas of law and philosophy, with a voluminous body of jurisprudence and academic theory addressing how rules limiting speech should be crafted.[5]

Today, this entire school of thought has been upended, as the practical functions of content regulation are being increasingly handed over to an industry which is not only grossly unprepared to handle the subtleties and technical challenges associated with defining the contours of acceptable speech on a global scale,[6] but has, as far as possible, resisted taking responsibility for this function.[7]

The origins of this dynamic, and the expanding privatization of content regulatory functions which have traditionally been performed by governments, lie, ironically, in the intermediary liability protections that date back to the earliest days of the commercial internet. Policymakers realized that the commercial and social potential of this new medium could best be realized if service providers were protected against being directly liable for the words of their users, spurring legal protections such as Section 230 of the *Communications Decency Act* in the United States.[8] These protections are what allowed scalability of the kind achieved by Facebook, Twitter, and other giant online platforms.

Having been allowed to grow without an expectation of closely policing their users, many of the world's biggest tech firms are built on business models which make it very difficult to control how their products are used. Now, governments around the world are increasingly complaining about the speech that emanates from these platforms and demanding that the companies assume greater responsibility for addressing objectionable content. In some cases, these demands involve fairly well recognized categories of harmful material, such as hate speech or child abuse imagery. Other examples involve content which is outlawed locally, but whose

---

[4] U.N. Human Rights Comm., *General Comment 34,* at art. 19, *Freedoms of Opinion and Expression*, U.N. Doc. CCPR/C/GC/34 (2011), https://www2.ohchr.org/english/bodies/hrc/docs/gc34.pdf.

[5] Kent Roach & David Schneiderman. *Freedom of Expression in Canada*, 61 SUP. CT. L. REV. (2d) (2013).

[6] *See, e.g.*, the persistent failure to apply fair use doctrine to alleged cases of copyright infringement: Leron Solomon, *Fair Users or Content Abusers? The Automatic Flagging of Non-Infringing Videos by Content ID on YouTube*, 44 HOFSTRA L. REV. 237 (2015).

[7] Even today, major platforms' default is to ask for regulation in this space: *See* Mark Zuckerberg, *Mark Zuckerberg: The Internet needs new rules. Let's start in these four areas*, WASH. POST (Mar. 30, 2019), https://www.washingtonpost.com/opinions/mark-zuckerberg-the-internet-needs-new-rules-lets-start-in-these-four-areas/2019/03/29/9e6f0504-521a-11e9-a3f7-78b7525a8d5f_story.html.

[8] 47 U.S.C. § 230 (1996).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

prohibition runs counter to global freedom of expression standards, such as risqué photos of the King of Thailand,[9] or material deemed insulting to Mustafa Kemal Atatürk, Turkey's founder.[10] In still other instances, governmental take-down requests are not grounded in any legal standard at all, but rather on the platforms' own independently drafted moderation standards.[11] These demands are backed by a variety of coercive measures, which depend on the character of the government, as well as the amount of leverage it holds over the intermediaries. However, the end result is a "privatized" system of content control, which is driven by government pressure, but which is also operated and enforced by the tech companies, in many cases sidestepping key judicial or constitutional safeguards that were developed to prevent the abusive application of content restrictions.

In response to this dynamic, the biggest global platforms have taken steps to rethink their approaches to content moderation. Facebook has taken the most decisive moves on this front, through its creation of a wholly independent Oversight Board.[12] However, there are broader calls for platforms to ground their approaches to moderation in independent standards of what constitutes harmful speech, rather than relying solely on their own internal considerations of the character of content that they want to host.[13] One natural focal point of discussions has been to use international freedom of expression standards as a guidepost for assessing what content should be permitted.[14] However, there are significant conceptual challenges to this approach, primarily since international freedom of expression standards were designed to apply to the actions of governments, rather than private sector actors. Nonetheless, a growing number of platforms seem to be embracing this idea to varying degrees. However, there is, as of yet, no agreed upon methodology for how

[9] Andrew Griffin, *Facebook Blocks Video of Thailand's King Wearing a Crop Top*, THE INDEPENDENT (May 11, 2017), https://www.independent.co.uk/life-style/gadgets-and-tech/news/facebook-thailand-king-thai-video-crop-top-bodindradebayavarangkun-maha-vajiralongkorn-a7729886.html.

[10] *Zuckerberg Notes Turkey's Defamation Laws over Ataturk as Facebook Updates Rules*, HURRIYET DAILY NEWS (Mar. 16, 2015), www.hurriyetdailynews.com/zuckerberg-notes-turkeys-defamation-laws-over-ataturk-as-facebook-updates-rules-79771.

[11] Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114, 177–78 (2018).

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

[13] *See, e.g.*, DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET (Colum. Glob. Rep. 2019).

[14] U.N. Human Rights Council, Comm'n on Human Rights, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 58, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018), https://undocs.org/A/HRC/38/35.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

exactly freedom of expression principles should be applied to the burgeoning field of "platform law."

This Article presents a preliminary framework for applying foundational freedom of expression standards to the context of private sector platforms. The first section of this Article traces the evolution of content moderation at private sector platforms, beginning with early light-touch approaches grounded in American First Amendment philosophy, and continuing through the adoption of more extensive moderation structures in response to growing global enforcement demands. This Article discusses both the rise in formal take-down requests, and of informal "jawboning" campaigns, which have led major platforms to become more aggressive and proactive in their enforcement.[15] This section concludes by discussing why the status quo is disfavored by virtually all stakeholders involved, culminating in a gradual move towards recognizing international freedom of expression standards as a core lodestar in developing and applying content restrictions. The second section introduces freedom of expression as an international human right, including the key sources of international standards that are used to interpret and apply this idea to practical regulatory challenges. The section also introduces core conceptual challenges to applying freedom of expression in the context of online platforms, namely in the gap between their role and that of government. The third section presents an avenue forward, and models how the three-part test, which lies at the core of understandings of freedom of expression as a human right, could be applied to platforms' moderation functions.

## I. THE ORIGINS OF "PLATFORM LAW"

### *A. The Early Days of First Amendment Dominance*

Although specific attitudes varied from company to company, the early years of tech platforms' ascendancy were marked by a fairly hands-off approach to content moderation. This was, in some cases, underpinned by a strong libertarian bend with regard to user speech. Dick Costolo, a former CEO of Twitter, famously described the company as being "the free speech wing of the free speech party."[16] Similarly, Reddit's then-CEO, Yishan Wong, said, in a post to the site's users: "We uphold the

[15] *See, e.g.*, Ángel Díaz, *Global Internet Forum to Counter Terrorism's 'Transparency Report' Raises More Questions Than Answers*, JUST SECURITY (Sept. 25, 2019), https://www.justsecurity.org/66298/gifct-transparency-report-raises-more-questions-than-answers/ (an initiative to target and remove "terrorist and violent extremist content" which was created under significant pressure from global governments).

[16] Emma Barnett, *Twitter Chief: We Will Protect Our Users from Government*, THE TELEGRAPH (Oct. 18, 2011, 10:23 AM), www.telegraph.co.uk/technology/twitter/8833526/Twitter-chief-We-will-protect-our-users-from-Government.html.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

ideal of free speech on Reddit as much as possible not because we are legally bound to, but because we believe that you—the user—has the right to choose between right and wrong, good and evil, and that it is your responsibility to do so."[17] More practical support for this position also flowed from the strong intermediary liability protections that these United States-based companies enjoyed as a result of Section 230 of *Communications Decency Act*, which allowed companies to avoid legal responsibility for the words of their users.[18] Similarly, in the European Union, the *Electronic Commerce Directive* protects intermediaries from liability so long as they act as a "mere conduit," "cache," or "host" of user expression.[19] In general, without a compelling legal requirement to police user content, the companies sought to avoid the trouble and expense of closely monitoring their customers.

There were exceptions to this relatively laissez-faire approach. In particular, complaints about the use of platforms as a vector for violating intellectual property rights led to some of the earliest moves towards widespread, automated content monitoring.[20] Some of these interventions were the result of legal requirements, such as the notice-and-take-down programs mandated by the *Digital Millennium Copyright Act*.[21] However, other moves to crackdown on intellectual property violations were an attempt to get ahead of future liability, or another round of tighter regulatory rules. In response to recurring complaints about infringement, including a high-profile lawsuit from Viacom,[22] in 2007 Google launched its Content Verification Program.[23] The Content Verification Program was subsequently rechristened as Content ID, a program to automatically screen uploaded content before it is posted against a database of protected files.[24] In November 2018, the company claimed to have spent over $100 million on developing and implementing

---

[17] Yishan Wong, *Every Man Is Responsible for His Own Soul*, REDDIT (Sept. 6, 2014), https://redditblog.com/2014/09/06/every-man-is-responsible-for-his-own-soul/.

[18] 47 U.S.C. § 230 (1996).

[19] Directive 2000/31 of the European Parliament and of the Council of 8 June 2000 on Certain Legal Aspects of Information Society Services, in Particular Electronic Commerce, in the Internal Market, 2000 O.J. (L 178) I (EC), https://eur-lex.europa.eu/legal-content/EN/ALL/?uri=CELEX:32000L0031.

[20] Taylor B. Bartholomew, *The Death of Fair Use in Cyberspace: YouTube and the Problem with Content ID*, 13 DUKE TECH. L. REV. 66, 71 (2015).

[21] Digital Millennium Copyright Act, Pub. L. No. 105-304, 112 Stat. 2860 (Oct. 28, 1998) 17 U.S.C. § 1201 *et seq.*

[22] Viacom Int'l Inc. v. YouTube, Inc., 718 F. Supp. 2d 514 (S.D.N.Y. 2010).

[23] Catherine Byni & Soraya Chemaly, *The Secret Rules of the Internet*, THE VERGE (Apr. 13, 2016), https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech.

[24] *How Content ID Works*, YOUTUBE (last accessed Apr. 7, 2020), https://support.google.com/youtube/answer/2797370?hl=en.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

Content ID.[25] Facebook has its own version of screening software, Rights Manager, though this was not introduced until 2016.[26]

Child abuse imagery was another area where platforms took strong proactive action. In 2009, Microsoft, in collaboration with Dartmouth College, developed PhotoDNA, a technology that screens for known images of child exploitation, as identified by law enforcement, particularly the National Center for Missing & Exploited Children.[27] This same technology is now also used by Facebook,[28] Twitter,[29] and Google,[30] with a common database of proscribed images. In September 2018, Google announced that its technology had further evolved, and was now capable of identifying new images of child abuse, as opposed to merely targeting previously confirmed images.[31]

However, apart from these exceptions, as well as certain efforts to combat spam and malware, early approaches to content moderation at the major platforms were carried out in a relatively unstructured and ad hoc basis.[32] Through the 1990s, early platforms essentially saw themselves as software companies, with speech implications as a mere "secondary effect of their primary business."[33] While platforms included various vaguely worded guidelines for acceptable content as part of their terms of service agreements, the practical enforcement of these policies was

---

[25] *How Google Fights Piracy*, GOOGLE 12 (Nov. 2018), https://storage.googleapis.com/gweb-uniblog-publish-prod/documents/How_Google_Fights_Piracy_2018.pdf.

[26] Analisa Tamayo Keef & Lior Ben-Kereth, *Facebook for Media: Introducing Rights Manager*, FACEBOOK (Apr. 12, 2016), https://www.facebook.com/facebookmedia/blog/introducing-rights-manager.

[27] *How Does PhotoDNA Technology Work?*, MICROSOFT (2019), https://www.microsoft.com/en-us/photodna.

[28] Facebook Security, *Want to know how Facebook uses photoDNA? Read a recent blog post by the head of our Safety Team*, FACEBOOK (Aug. 10, 2011), https://www.facebook.com/security/posts/want-to-know-how-facebook-uses-photodna-read-a-recent-blog-post-by-the-head-of-o/234737053237453/.

[29] Charles Arthur, *Twitter to Introduce PhotoDNA System to Block Child Abuse Images*, GUARDIAN (July 22, 2013), https://www.theguardian.com/technology/2013/jul/22/twitter-photodna-child-abuse.

[30] Rich McCormick, *Google Scans Everyone's Email for Child Porn, and it Just Got a Man Arrested*, THE VERGE (Aug. 5, 2014), https://www.theverge.com/2014/8/5/5970141/how-google-scans-your-gmail-for-child-porn.

[31] Nikola Todorovic & Abhi Chaudhuri, *Using AI to Help Organizations Detect and Report Child Sexual Abuse Material Online*, THE KEYWORD (Sept. 3, 2018), https://www.blog.google/around-the-globe/google-europe/using-ai-help-organizations-detect-and-report-child-sexual-abuse-material-online/.

[32] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1618 (2018).

[33] *Id.*



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

heavily influenced by traditional American understandings of freedom of speech.[34] The platforms' early struggles in coming to grips with these challenges were exacerbated by the fact that the decision-makers had virtually no experience in global freedom of expression norms, or the local human rights situation across the many countries where the platforms were active.[35]

### *B. The First Amendment Meets the Global Market*

Since the earliest days of the commercial internet, the fact that so much of the infrastructure that facilitates global online communications is based in the United States has been a source of controversy.[36] These tensions inevitably impact the operations of American tech firms doing business abroad, particularly where the local political environment is substantially different from the liberal-democratic model.[37] In general, when a company does business in a particular country, there are reasonable expectations that it should abide by local law and customs. However, it is obviously not a good look to be seen collaborating with human rights abuses, including the repression of dissenting speech.[38]

Companies can always choose not to do business in places which demand ethical compromises. However, in addition to the commercial cost that withdrawing from a country entails, this calculus is made more challenging by the fact that human rights determinations are rarely black and white. No country has a perfectly clean human rights record. "Bad" countries can make perfectly reasonable demands, and "good" countries can make abusive ones. It would be silly to expect Facebook to reject a notification of child sexual abuse material just because it was delivered by the Russian government, and it is similarly problematic for a company to blindly acquiesce to any request emanating from American or Canadian law enforcement.

As American platforms became global platforms, they inevitably faced growing challenges with regard to differing understandings of what types of content should be prohibited. With YouTube, for example, early requests from Thailand and Turkey in 2006 and 2007 were the first time they had to puzzle through their approach to countries whose understandings of freedom of expression differed vastly

---

[34] *Id.* at 1621.

[35] *Id.* at 1618–19.

[36] *See generally* MILTON L. MUELLER, RULING THE ROOT (2002).

[37] *See, e.g.*, Joseph Kahn, *Yahoo helped Chinese to prosecute journalist*, N.Y. TIMES (Sept. 8, 2005), http://www.nytimes.com/2005/09/07/business/worldbusiness/07iht-yahoo.html.

[38] Ronald J. Deibert, *The Road to Digital Unfreedom: Three Painful Truths About Social Media*, 30 J. OF DEMOCRACY 25, 35 (2019).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

from First Amendment values.[39] In some cases, content requests touched on highly controversial subjects, such as with Facebook's decision to remove pages linked to the Peace and Democracy Party (BDP), which was then Turkey's largest pro-Kurdish political party.[40] Facebook stated that the removals were linked to content supportive of the Kurdistan Workers' Party (PKK), which it found to be a violation of its prohibition on expressing support for internationally-recognized illegal terrorist organizations.[41] But while some governments have labeled the PKK as terrorists,[42] not all have, and the question of whether they are a terrorist group or a legitimate independence movement is an intensely political one.[43]

Moderation questions around Palestinian content tap into an even more globally polarized debate.[44] Facebook, Twitter and Google have faced enormous pressure from the Israeli government to remove material which it considers as inciting violence or anti-Semitism.[45] The Palestinian Center for Development and Media Freedoms (MADA), a leading civil society organization which works to promote freedom of expression in Palestine, has complained that the resulting collaborations between the companies and the Israeli government have led to legitimate journalists and news organizations losing their accounts.[46] In the context of the limited media landscape in Palestine, deplatforming can be a crippling blow for a journalist or news outlet, and the systemic targeting of these accounts has a substantial impact on Palestinians' freedom of expression and access to information, as well as their ability to communicate local perspectives globally.[47]

---

[39] Klonick, *supra* note 32, at 1623.

[40] Jacob Resneck, *Facebook Censorship of Pro-Kurdish Political Party*, DEUTSCHE WELLE (Nov. 2, 2013), https://www.dw.com/en/facebook-censorship-of-pro-kurdish-political-party/a-17199752.

[41] *Id.*

[42] U.S. Dep't of State, *Foreign Terrorist Organizations*, https://www.state.gov/foreign-terrorist-organizations/.

[43] *See, e.g.*, Mehmet Alper Sozer & Kamil Yilmaz, *The PKK and its evolution in Britain (1984–present)*, 31 TERRORISM & POL. VIOLENCE 185 (2019).

[44] Palestinian Center for Development and Media Freedoms, *Social Media . . . A New Venue to Censor and Prosecute Journalists* (Oct. 20, 2016), https://www.madacenter.org/files/image/editor/FBviolationsE.pdf.

[45] Sue Surkes, *Shaked: Facebook, Twitter removing 70% of 'harmful' posts*, TIMES OF ISRAEL (June 7, 2016), http://www.timesofisrael.com/70-of-harmful-facebook-twitter-posts-said-removed/.

[46] *Social Media . . . A New Venue to Censor and Prosecute Journalists*, Palestinian Center for Development and Media Freedoms, *supra* note 44.

[47] *No News is Good News: Abuses against Journalists by Palestinian Security Forces*, HUMAN RIGHTS WATCH (Apr. 6, 2011), https://www.hrw.org/report/2011/04/06/no-news-good-news/abuses-against-journalists-palestinian-security-forces.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

The above cases are illustrative of how platforms are increasingly finding themselves in the middle of politically charged international debates, and essentially forced to choose one side or the other. Often, the calculus underlying this decisionmaking is unclear, and policies appear reactionary and driven by external pressure. In 2019, the United States Government made a decision to officially designate Iran's Islamic Revolutionary Guard Corps (IRGC) as a terrorist organization, the first time this label has been applied to a foreign governmental apparatus.[48] The following day, Facebook began to delete accounts related to the organization and its officers.[49] Regardless of one's opinion of the decision to designate a branch of a foreign military as terrorists, this is fundamentally a dispute between the governments of the United States and Iran. While it may not be surprising that Facebook followed the American position rather than the Iranian one, it is nonetheless significant that the company is overtly choosing sides in a state-to-state conflict.

The relationship between State pressure and platforms' enforcement is further complicated by the fact that requests from governments are increasingly based not on violations of local law, but on platforms' own content policies.[50] Part of the reason for this is because of the vague terms in which platforms craft their categories of prohibited speech.[51] This definition ensures that the platform maintains relative flexibility to act against content it finds problematic, but it also turns it into a vulnerable target for governments seeking to remove content that they do not have the legal authority to target directly.[52] A number of States, particularly in Europe, have constituted specialized referral agencies which are tasked with monitoring social media and filing take-down requests based on perceived terms of service violations.[53] According to its 2017 transparency reporting, the European Union's Internet Referral Unit enjoyed a 92% "success rate" for platform removal requests

[48] Robert Wright, *Why Is Facebook Abetting Trump's Reckless Foreign Policy?*, WIRED (May 7, 2019), https://www.wired.com/story/why-is-facebook-abetting-trumps-reckless-foreign-policy/.

[49] *Id.*

[50] Chang, *supra* note 11.

[51] *The Twitter Rules*, TWITTER (last visited Apr. 7, 2020), https://help.twitter.com/en/rules-and-policies/twitter-rules.

[52] U.N. Human Rights Council, Comm'n on Human Rights, *Report of the Special Rapporteur on the Promotion and Protection of Human and Fundamental Freedoms while Countering Terrorism, Martin Scheinin*, U.N. Doc. A/HRC/16/51 (Dec. 22, 2010), https://undocs.org/A/HRC/16/51.

[53] Matthias Monroy, *German Police launches "National Internet Referral Unit,"* SEC. ARCHITECTURES AND POLICE COLLABORATION IN THE EU (Mar. 24, 2020, 11:04 PM), https://digit.site36.net/2019/04/24/german-police-launches-national-internet-referral-unit/.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

dealing with "terrorist content."[54] The same Report makes clear that because the material is assessed against the terms of service of the relevant platform, and because the ultimate decision to remove the material is in the hands of the platform, their referral activity "does not constitute an enforceable act" by the agency.[55]

### *C. "Self-regulation" and Moves Towards Proactive Enforcement*

Beyond relatively targeted take-down efforts, platforms have faced increasing pressure to introduce more proactive measures at targeting content which governments find objectionable. This process is generally known as "jawboning," or moral suasion, whereby platforms are pressured through threats of regulation to shift their broader approach to moderating content in order to bring it into line with categories that governments might seek to target.[56] In other words, rather than initiating a formal take-down process against a particular item, senior officials will loudly complain about a particular type of content, including pointing to specific examples, and declare that costly and restrictive laws will be considered if the issue is not resolved to the government's satisfaction.

There are a number of reasons why jawboning is effective. First, platforms may view an independently managed policy-shift as cheaper and less unpredictable than having to comply with new, binding rules.[57] Jawboning also neatly sidesteps constitutional challenges that might stand in the way of enacting new laws, as well as any political resistance if the new rules are unpopular or controversial. Governments who face resistance, either from the public or internally, to new legislation may be able to successfully bluff that the laws are just over the horizon.

However, while there can be benefits to governments employing jawboning as a regulatory strategy to push for greater private sector responsibility,[58] in the context of restrictions on expression this tactic raises concerns. First, it removes the opportunity to constitutionally challenge new rules, or to seek judicial clarity regarding the scope of speech which is now prohibited. Indeed, the potential for jawboning to provide an effective avenue to end-run constitutional limits can be

---

[54] EUROPOL, EU INTERNET REFERRAL UNIT, TRANSPARENCY REPORT 5 (2017), https://www.europol.europa.eu/publications-documents/eu-internet-referral-unit-transparency-report-2017.

[55] *Id.* at 4.

[56] Derek E. Bambauer, *Against Jawboning*, 100 MINN. L. REV. 50, 57 (2015).

[57] Daphne Keller, *Who Do You Sue? State and Platform Hybrid Power Over Online Speech* 2, LAWFARE (Jan. 29, 2019, 2:37 PM), https://www.lawfareblog.com/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech.

[58] *See, e.g.*, Ronald W. Cotterill, *Jawboning Cereal: The Campaign to Lower Cereal Prices*, 15 AGRIBUSINESS 197 (1998).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

viewed as a substantial challenge to democratic norms.[59] Second, if the new restrictions are unpopular, it is far more difficult for the public to express their displeasure at the ballot box, since it can be challenging to point to a direct causal connection between government statements and the changes that are ultimately instituted. While this dynamic is characteristic of all jawboning campaigns, it is particularly problematic given the political nature of speech restrictions, and the importance of freedom of speech to the political process.

Among the most prominent results of governmental pressure campaigns has been the creation of the Global Internet Forum to Counter Terrorism (GIFCT), an industry-led self-regulatory initiative that includes Facebook, Microsoft, Twitter, and YouTube.[60] GIFCT works by supporting the development of machine-learning algorithms to catch extremist content, as well as through the development of a shared hash database, which participating companies use to flag problematic content for removal.[61] Although the collaboration has been particularly effective at combating the spread of ISIS and al-Qaeda-related material, civil society voices have criticized its lack of transparency, as well as the way content moderation functions are being centralized and harmonized.[62] Part of the reason why content removal by private sector actors has traditionally been seen as less intrusive than State blocking or censoring mechanisms is due to the diversity of the platforms available, allowing users to seek a new audience elsewhere if a particular platform rejects them. If major platforms begin to consolidate their standards, private sector decisions may take on the gravity of a global ban.

Even more concerning has been the drift from demanding action to curtail hate speech and incitement to violence, which are legitimate areas of State regulation, to pushing for more intensive efforts to combat misinformation, or "fake news."[63] For example, in May 2019 Canada's Prime Minister Justin Trudeau announced as part

---

[59] Michael Karanicolas, *Subverting Democracy to Save Democracy: Canada's Extra-Constitutional Approaches to Battling 'Fake News*,' 17 CAN. J.L. & TECH. 201 (2019).

[60] *About Our Mission*, GLOBAL INTERNET FORUM TO COUNTER TERRORISM (last visited Mar. 27, 2020), https://www.gifct.org/about/.

[61] Press Release, Global Internet Forum to Counter Terrorism, Facebook, Microsoft, Twitter and YouTube Announce Formation of the Global Internet Forum to Counter Terrorism (June 26, 2017), https://gifct.org/press/facebook-microsoft-twitter-and-youtube-announce-formation-global-internet-forum-counter-terrorism/.

[62] Opinion, Emma Llansó, *Platforms Want Centralized Censorship. That Should Scare You*, WIRED (Apr. 18, 2019), https://www.wired.com/story/platforms-centralized-censorship/.

[63] Danielle Keats Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1039 (2018).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

of the country's "Digital Charter" that platforms which failed to adequately take action against misinformation would face "meaningful financial consequences."[64]

In contrast to hate speech and incitement, laws prohibiting misinformation are generally not acceptable according to international freedom of expression standards. This was reiterated in the 2017 Joint Statement by the Special Mandates on Freedom of Expression, which specifically addressed responses to misinformation:

> *Stressing* that the human right to impart information and ideas is not limited to "correct" statements, that the right also protects information and ideas that may shock, offend and disturb, and that prohibitions on disinformation may violate international human rights standards, while, at the same time, this does not justify the dissemination of knowingly or recklessly false statements by official or State actors;
>
> . . .
>
> 2. Standards on Disinformation and Propaganda:
>
> a. General prohibitions on the dissemination of information based on vague and ambiguous ideas, including "false news" or "non-objective information," are incompatible with international standards for restrictions on freedom of expression, as set out in paragraph 1(a), and should be abolished.[65]

There are a number of reasons why misinformation laws are problematic, including that they essentially grant authorities the power to dictate a single, correct version of "the truth," and to target voices which deviate from that agreed narrative. While it is easy to point to harmful kinds of misinformation, such as, for example, the distribution of fictitious theories about the dangers of vaccination,[66] it is practically impossible to craft a legal prohibition against misinformation which would not lend itself to abuse in targeting more benign forms of dissent. Where laws

[64] *Trudeau Warns of Meaningful Financial Consequences for Social Media Giants that Don't Combat Hate Speech*, CBC (May 16, 2019, 8:49 AM), https://www.cbc.ca/news/politics/digital-charter-trudeau-1.5138194.

[65] Office of the Special Rapporteur for Freedom of Expression, *Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda*, ORG. AM. STS. (Mar. 3, 2017), www.oas.org/en/iachr/expression/showarticle.asp?artID=1056&lID=1.

[66] Ayelet Evrony & Arthur Caplan, *The Overlooked Dangers of Anti-Vaccination Groups' Social Media Presence*, 13:6 HUM. VACCIN. & IMMUNOTHER. 147 (2017).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

against "fake news" are in force, they are often abused to target opposition politicians, journalists, and civil society.[67]

### *D. Pressure from All Sides*

The pressures described here have left platforms between a rock and a hard place when it comes to content moderation. While most major platforms have been loath to accept full responsibility for the content that they host or otherwise facilitate,[68] it seems manifestly clear that the current status quo is untenable, given the dissatisfaction of all major stakeholders, and growing tensions permeating the system.

Governments are pressuring major intermediaries to do more and appear increasingly unsatisfied with the perceived inconsistency and unreliability in private content moderation structures.[69] In parallel to this general dissatisfaction, governmental demands have escalated over the past few years. In part, this may be attributed to a kind of "ratchet effect,"[70] where all governments demand that their requests be treated with the highest level of priority and urgency that platforms accord to other governments. As a result, acquiescing to demands from one government can lead to escalating demands from many others. Since the European Commission's *Code of Conduct on Countering Illegal Hate Speech Online* first imposed a 24-hour turnaround time for removing certain content, that standard has been followed by other governments.[71] More recently, proposals from the European Parliament could push the deadline for responding to "terrorist content" notifications

---

[67] Daniel Funke, *A Guide to Anti-misinformation Actions Around the World*, THE POYNTER INSTITUTE (Oct. 31 2018), https://www.poynter.org/fact-checking/2018/a-guide-to-anti-misinformation-actions-around-the-world/.

[68] Yishan Wong, *Every Man Is Responsible for His Own Soul*, REDDIT (Sept. 6, 2014), https://redditblog.com/2014/09/06/every-man-is-responsible-for-his-own-soul/.

[69] Makena Kelly, *Facebook, YouTube, and Others Asked to Brief Congress on New Zealand Shooting Response*, THE VERGE (Mar. 19, 2019), https://www.theverge.com/2019/3/19/18273257/facebook-youtube-microsoft-twitter-congress-zealand-shooting-response.

[70] What is Ratchet Effect?, THE LAW DICTIONARY (last visited Apr. 7, 2020), https://thelawdictionary.org/ratchet-effect/.

[71] *See, e.g.*, Javier Pallero, *Honduras: New Bill Threatens to Curb Online Speech*, ACCESS NOW (Feb. 12, 2018), https://www.accessnow.org/honduras-new-bill-threatens-curb-online-speech/ (discussing Honduras' proposed cybersecurity bill); *see also* Paris Martineau, *India's Plan to Curb Hate Speech Could Mean More Censorship*, WIRED (Jan. 10, 2019), https://www.wired.com/story/indias-plan-curb-hate-speech-mean-more-censorship/?verso=true (discussing the fact that India is considering implementing a 24-hour turnaround time for removal of illegal content as part of reforms to their safe harbor rules).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

to just one hour, which would make meaningful assessment of the material under review practically impossible.[72]

The platforms have expressed deep discomfort with the level of responsibility and power they wield over the global expressive discourse.[73] Nicole Wong, a former lawyer at Google, admitted her own difficulty in trying to deal with "the norms of behavior when what's appropriate is constantly reiterated. If you layer over all of that the technology change and the cultural, racial, national, [and] global perspectives, it's all just changing dramatically fast. It's enormously difficult to figure out those norms, let alone create policy to reflect them."[74]

This dynamic is further complicated by the commercial pressures that platforms face in hosting controversial speech, since the companies mainly rely on advertising to draw in revenue. In March 2017, a number of brands began to pull advertising from YouTube after discovering their ads were being posted next to extremist content.[75] A similar exodus of advertisers took place in 2019, after it was discovered that users were leaving sexually suggestive comments on videos of young girls, and potentially utilizing the site's recommendation algorithm to find similar videos.[76] YouTube's actions in responding to these issues were not particularly controversial.[77] But the episodes illustrate a potential tension between advertisers, who may be reluctant to have their products associated with polarizing ideas, and users, who view platforms as their primary mechanism for self-expression and political engagement, and may dislike moderation structures that substantially narrow the boundaries of acceptable content to that which is universally viewed as uncontroversial. This tension exacerbates a misaligned incentive structure, noted by

---

[72] *Commission Proposal for a Regulation of the European Parliament and of the Council on Preventing the Dissemination of Terrorist Content Online*, COM (2018) 640 final (Sept. 12, 2018).

[73] *See, e.g.*, Kara Swisher & Kurt Wagner, *Here's the transcript of Recode's interview with Facebook CEO Mark Zuckerberg about the Cambridge Analytica controversy and more*, VOX (Mar. 22, 2018), https://www.vox.com/2018/3/22/17150814/transcript-interview-facebook-mark-zuckerberg-cambridge-analytica-controversy.

[74] Klonick, *supra* note 32, at 1628.

[75] Lara O'Reilly, *The Real Motivations Behind the Growing YouTube Advertiser Boycott*, BUS. INSIDER (Mar. 22, 2017), https://www.businessinsider.my/why-advertisers-are-pulling-spend-from-youtube-2017-3/.

[76] Jim Waterson, *Fortnite Maker Pulls Ads over YouTube 'Paedophile Ring' Claims*, THE GUARDIAN (Feb. 21, 2019), https://www.theguardian.com/technology/2019/feb/21/fortnite-maker-pulls-ads-over-youtube-paedophile-ring-claims.

[77] Alex Hern, *YouTube turns off comments on videos of children amid child safety fears*, THE GUARDIAN (Feb. 28, 2019), https://www.theguardian.com/society/2019/feb/28/youtube-turns-off-comments-on-videos-of-children-amid-child-safety-fears.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

Jack Balkin among others, between platforms as media companies and platforms as surveillance companies, with the latter representing their core business model.[78]

The unreliable and reactionary nature of the current system undermines the interests of users, advertisers, governments, and the platforms themselves. More broadly, the past few years have seen a public backlash against the unprecedented power that platforms wield. Since the 2016 scandals over the use of online platforms to interfere in both the United States presidential election and the United Kingdom's Brexit vote,[79] there have been a near constant stream of negative media stories about content-moderation failures at major tech firms. These include allegations that moderation standards discriminate against LGBTQ users[80] as well as other marginalized groups,[81] that recommendation algorithms are helping to spread anti-vaccination misinformation,[82] and that YouTube's algorithms were removing evidence of Syrian chemical weapons attacks which had been uploaded by journalists or witnesses on the ground, and which in some instances may not have been backed up elsewhere.[83] The volume of stories, and the growing public sentiment against major platforms, has led to prominent calls for Facebook, in particular, to be broken up,[84] or even nationalized.[85]

---

[78] Jack M. Balkin, *How to Regulate (and Not Regulate) Social Media*, SSRN (Nov. 8, 2019), https://ssrn.com/abstract=3484114.

[79] Henry Farrell & Bruce Schneier, *Common-Knowledge Attacks on Democracy*, BERKMAN KLEIN CTR. FOR INTERNET & SOC'Y, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3273111.

[80] Samantha Allen, *Social Media Giants Have a Big LGBT Problem. Can They Solve It?*, THE DAILY BEAST (Oct. 12, 2018), https://www.thedailybeast.com/social-media-giants-have-a-big-lgbt-problem-can-they-solve-it.

[81] Tracy Jan & Elizabeth Dwoskin, *A White man called her kids the n-word Facebook stopped her from sharing it*, WASH. POST (July 31, 2017), https://www.washingtonpost.com/business/economy/ for-facebook-erasing-hate-speech-proves-a-daunting-challenge/2017/07/31/922d9bc6-6e3b-11e7-9c15-177740635e83_story.html.

[82] Julia Carrie Wong, *How Facebook and YouTube Help Spread Anti-Vaxxer Propaganda*, THE GUARDIAN (Feb. 1, 2019), https://www.theguardian.com/media/2019/feb/01/facebook-youtube-anti-vaccination-misinformation-social-media.

[83] Kate O'Flaherty, *YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks*, WIRED (June 26, 2018), https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video.

[84] Chris Hughes, *It's Time to Break Up Facebook*, N.Y. TIMES (May 9, 2019), https://www.nytimes.com/2019/05/09/opinion/sunday/chris-hughes-facebook-zuckerberg.html.

[85] Blayne Haggart, *Why Not Nationalize Facebook?*, NAT'L POST (Mar. 31, 2018), https://nationalpost.com/pmn/news-pmn/why-not-nationalize-facebook.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

### *E. Searching for an Exit Strategy*

In considering solutions to this problem, an increasing number of voices have pointed to international human rights rules as the key lodestar to developing global solutions to moderating online content. David Kaye, the United Nations Special Rapporteur on the promotion and protection of the right to freedom of opinion and expression has been particularly vocal as an advocate for an approach guided by human rights norms:

> Companies should recognize that the authoritative global standard for ensuring freedom of expression on their platforms is human rights law, not the varying laws of States or their own private interests, and they should re-evaluate their content standards accordingly. Human rights law gives companies the tools to articulate and develop policies and processes that respect democratic norms and counter authoritarian demands. This approach begins with rules rooted in rights, continues with rigorous human rights impact assessments for product and policy development, and moves through operations with ongoing assessment, reassessment and meaningful public and civil society consultation.[86]

Increasingly, it appears as though platforms are receptive to this idea. The most significant announcement in this regard has been Facebook's decision to create an independent Oversight Board, which Mark Zuckerberg, in announcing the initiative, compared to the platform having its own "Supreme Court."[87] In January 2020, Facebook released its proposed bylaws for the Oversight Board, which included a commitment to implement their decisions in a way which is "guided by relevant human rights principles."[88] It also named the Oversight Board's first Director as Thomas Hughes, who had previously served as the Executive Director of Article 19, a global human rights organization.[89] The choice of such a prominent figure from the human rights community is strongly suggestive of an intent for international human

---

[86] U.N. Human Rights Council, Comm'n on Human Rights, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, ¶ 70, U.N. Doc. A/HRC/38/35 (Apr. 6, 2018), https://undocs.org/A/HRC/38/35.

[87] Ezra Klein, *Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next* (Apr. 2, 2018), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.

[88] FACEBOOK, OVERSIGHT BOARD BYLAWS (Jan. 2020), https://about.fb.com/wp-content/uploads/2020/01/Bylaws_v6.pdf.

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



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

rights standards, particularly around freedom of expression, to play a major role in the Board's decision-making.

While Facebook has taken the most significant steps towards recognizing the role of international freedom of expression standards in guiding its decision-making, they are not alone in charting this path. Twitter's site rules, for example, feature statements pointing to the role of freedom of expression as an international human right in its decision-making, including a specific reference to the *European Convention on Human Rights*.[90] Likewise, the *Global Network Initiative (GNI)*, which counts a number of prominent intermediaries including Facebook, Google, and Microsoft among its members, makes specific reference to the role of internationally recognized laws and standards for human rights, including the *International Covenant on Civil and Political Rights (ICCPR)*,[91] in developing their core standards.[92] The GNI principles include references to freedom of expression, though it is worth noting that this is specifically defined in terms of government restrictions on speech, and not the platforms' own content guidelines.[93] In early 2020, TikTok announced its own plans to constitute a "Content Advisory Council" to provide external input into its content moderation policies.[94] Though this process does not appear to be modelled on human rights norms, it is significant insofar as it reflects the company's sensitivity to the need to seek broader legitimacy around decisions in this space.

## II. UNDERSTANDING AND APPLYING FREEDOM OF EXPRESSION STANDARDS

### *A. Freedom of Expression as an International Human Right*

Freedom of expression is a core human right, which is protected under virtually every constitution in the world, along with all of the main international and regional human rights treaties. This includes, most notably, the *Universal Declaration of Human Rights*, which was adopted unanimously by the United Nations General

[90] *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 (last visited Apr. 10, 2020).

[91] International Covenant on Civil and Political Rights, Dec. 19, 1966, 999 U.N.T.S. 171.

[92] Global Network Initiative, GNI Principles on Freedom of Expression and Privacy 1 (2018), https://globalnetworkinitiative.org/gni-principles/.

[93] *Id.*

[94] Vanessa Pappas, *Introducing the TikTok Content Advisory Council*, TIKTOK (Mar. 18, 2020), https://newsroom.tiktok.com/en-us/introducing-the-tiktok-content-advisory-council.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

Assembly in 1948: "Everyone has the right to freedom of opinion and expression; this right includes freedom to hold opinions without interference and to seek, receive and impart information and ideas through any media and regardless of frontiers."[95]

The status of freedom of expression as a right is reinforced and expanded in Article 19 of the *International Covenant on Civil and Political Rights (ICCPR)*:

> 1. Everyone shall have the right to hold opinions without interference.
>
> 2. Everyone shall have the right to freedom of expression; this right shall include freedom to seek, receive and impart information and ideas of all kinds, regardless of frontiers, either orally, in writing or in print, in the form of art, or through any other media of his choice.
>
> 3. The exercise of the rights provided for in paragraph 2 of this article carries with it special duties and responsibilities. It may therefore be subject to certain restrictions, but these shall only be such as are provided by law and are necessary:
>
> (a) For respect of the rights or reputations of others;
>
> (b) For the protection of national security or of public order (ordre public), or of public health or morals.[96]

Similar language has been incorporated into a number of regional treaties, such as the *African Charter on Human and People's Rights*,[97] the *European Convention on Human Rights*,[98] and the *American Convention on Human Rights*.[99]

A core aspect of the right to freedom of expression, as framed in the ICCPR, is that restrictions on speech are required to comply with a strict three-part test: the restrictions must be "provided by law;" they may only be imposed for a legitimate purpose; and they must conform to strict tests of necessity and proportionality.

International human rights treaties, and in particular the *ICCPR*, have given rise to a robust body of international human rights standards on interpreting the right to

---

[95] G.A. Res. 217 (III) A, Universal Declaration of Human Rights (Dec. 10, 1948).

[96] International Covenant on Civil and Political Rights, *supra* note 91.

[97] African Charter on Human and Peoples' Rights, OAU Doc. CAB/LEG/67/3 rev. 5, 21 I.L.M. 58 (1982) (entered into force Oct. 21, 1986).

[98] Council of Europe, *European Convention for the Protection of Human Rights and Fundamental Freedoms*, as amended by Protocols Nos. 11 and 14 and supplemented by Protocols Nos. 1, 4, 6, 7, 12 and 13, Nov. 4, 1950, C.E.T.S. 5.

[99] American Convention on Human Rights, Nov. 22, 1969, 1144 U.N.T.S. 123, 9 I.L.M. 673 (entered into force July 18, 1978), art. 13.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

freedom of expression, beginning with the *General Comments* issued by the United Nations Human Rights Committee. The most recent of these addressing freedom of expression was *General Comment No. 34*, which was published in 2011.[100] It includes a thorough discussion on interpreting the three-part test, including that restrictions on speech must meet standards of clarity to be considered as having been "provided by law," such as being publicly accessible, being constructed with "sufficient precision to enable an individual to regulate his or her conduct accordingly," and limiting the degree of discretion related to their execution.[101] Partly this standard is related to fundamental notions of procedural fairness, but it is also meant to limit, as far as possible, any potential "chilling effect," where uncertainty about what is and is not permitted causes people to steer well clear of the line, and potentially even avoid controversial topics entirely. *General Comment No. 34* also further elaborates on the standard of "necessity," which incorporates a basic assessment of proportionality:

> Restrictions must not be overbroad. The Committee observed *General Comment No. 27* that "restrictive measures must conform to the principle of proportionality; they must be appropriate to achieve their protective function; they must be the least intrusive instrument amongst those which might achieve their protective function; they must be proportionate to the interest to be protected . . . . The principle of proportionality has to be respected not only in the law that frames the restrictions but also by the administrative and judicial authorities in applying the law." The principle of proportionality must also take account of the form of expression at issue as well as the means of its dissemination. For instance, the value placed by the Covenant upon uninhibited expression is particularly high in the circumstances of public debate in a democratic society concerning figures in the public and political domain.
>
> When a State party invokes a legitimate ground for restriction of freedom of expression, it must demonstrate in specific and individualized fashion the precise nature of the threat, and the necessity and proportionality of the specific action taken, in particular by establishing a direct and immediate connection between the expression and the threat. [references omitted][102]

[100] U.N. Human Rights Comm., *General Comment 34*, *supra* note 4.

[101] *Id.* ¶ 25.

[102] *Id.* ¶¶ 34–35.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

### *B. Interpreting and Applying Freedom of Expression*

There are a wide variety of sources for drilling down into the applicability of broad human rights concepts, like the three-part test, to specific challenges. These include decisions of the UN Human Rights Committee, as well as regional bodies such as the European Court of Human Rights and the Inter-American Court of Human Rights, and statements from special mandates appointed by standard setting organizations, such as the United Nations Special Rapporteur on Freedom of Opinion and Expression, and the Organization for Security and Co-operation in Europe Representative on Freedom of the Media.

In addition to appropriate treaty law, jurisprudence, and statements, scholarly treatment of freedom of expression can be another source for establishing international human rights standards, along with publications or statements from expert civil society groups specializing in this area (such as Article 19,[103] Human Rights Watch,[104] and the Centre for Law and Democracy[105]). Joint civil society statements or declarations can be particularly useful, as they may demonstrate a consensus among advocates working on these issues. For example, the *Necessary and Proportionate Principles on the Application of Human Rights to Communications Surveillance* have attracted an extremely broad base of signatories, reflecting widespread support among global human rights experts and practitioners.[106]

None of these sources are prescriptive, in terms of providing a specific, universally applicable formula for how speech should be governed. Nonetheless, taken together they provide a robust and extensive body of standards which are useful to inform regulatory approaches. In particular, they can assist in defining what an unacceptable restriction looks like, and identifying potential areas of concern with a given regulatory strategy based on this understanding.

---

[103] *Defining Defamation: Principles on Freedom of Expression and Protection of Reputation*, ART. 19 (Feb. 23, 2017), https://www.article19.org/resources/defining-defamation-principles-on-freedom-of-expression-and-protection-of-reputation/.

[104] *Promote Strong Encryption and Anonymity in the Digital Age*, HUMAN RIGHTS WATCH (June 17, 2015), https://www.hrw.org/news/2015/06/17/promote-strong-encryption-and-anonymity-digital-age-0.

[105] TOBY MENDEL, *Restricting Freedom of Expression: Standards and Principles*, CENTRE L. & DEMOC. (Mar. 2010), http://www.law-democracy.org/wp-content/uploads/2010/07/10.03.Paper-on-Restrictions-on-FOE.pdf [hereinafter MENDEL, *Restricting*].

[106] *International Principles on the Application of Human Rights to Communications Surveillance*, NECESSARY & PROPORTIONATE (May 2014), https://en.necessaryandproportionate.org/ (last visited Apr. 11, 2020).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

For example, with regard to hate speech or incitement, while specific national approaches to regulation differ, international human rights standards have coalesced around a number of key elements for how these laws should operate. The starting point is Article 19 of the ICCPR, along with Article 20(2), which mandates that "advocacy of national, racial or religious hatred that constitutes incitement to discrimination, hostility or violence shall be prohibited by law."[107] However, the text leaves open questions as to how one should craft an appropriate balance between this rule and the right to freedom of expression. Relevant case law is appropriate to tease out guiding principles, such as the requirement that prescribed speech should be intended to generate hatred, which is important to ensure that journalists or anti-racist advocacy groups who may reprint hateful statements to call attention to them are not caught under these rules.[108] Likewise, international courts have repeatedly found that assessing whether something is hate speech is an inherently contextual determination, based on whether the material was released under circumstances where it is likely to actually generate harm.[109] This means, for example, that an identical statement made in Myanmar and in Canada might be considered hate speech in the former but not in the latter, given the differing levels of underlying ethnic tensions which would be relevant to assessing its impact.

To take another example, while the *ICCPR* recognizes the legitimacy of rules to protect "the rights or reputations of others" (i.e., defamation), recent treatment of this issue has focused on the need to consider the proportionality of these restrictions carefully, including a preference for civil restrictions rather than criminal ones.[110]

Even in dealing with obscenity, an area of regulation that is especially tied to local cultures and traditions, it is possible to draw out some useful standards. In *General Comment No. 34*, the UN Human Rights Committee noted the importance of developing standards of morality in an inclusive and non-discriminatory manner:

> The Committee observed in *General Comment No. 22*, that "the concept of morals derives from many social, philosophical and religious traditions; consequently, limitations . . . for the purpose of protecting morals must be based on principles not deriving exclusively from a single tradition." Any such limitations must be

[107] International Covenant on Civil and Political Rights, *supra* note 91.

[108] Jersild v. Denmark, No. 15890/89, ECHR 33, 19 EHRR 1 (1994).

[109] Erbakan v. Turkey, App. No. 59405/00 (ECtHR, 6 July 2006); Prosecutor v. Nahimana, Case No. ICTR-99-52-T, Judgement and Sentence (Dec. 3, 2003).

[110] Abid Hussain, Freimut Duve & Santiago Canton, *International Mechanisms for Promoting Freedom of Expression Joint Declaration*, ORG. AM. STATES (Nov. 30, 2000), http://www.oas.org/en/iachr/expression/showarticle.asp?artID=142&lID=1.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

> understood in the light of universality of human rights and the principle of non-discrimination[111]

This last example is particularly illustrative of how international human rights standards present a set of values and characteristics that should inform the crafting and application of laws. Another way to think about this would be that international human rights standards define an appropriate spectrum for domestic regulatory approaches, as well as general values which should inform their development and application. While they typically are not prescriptive of exactly where on that spectrum a country should place itself, they anticipate that a regulatory approach should take place somewhere within these defined boundaries.

### *C. Challenges in Applying Freedom of Expression to the Context of Platforms*

International freedom of expression standards are generally understood to restrict the actions of governments, rather than private actors. This is not to suggest a complete divorce between *ICCPR* principles and the private sector. Molly Land has argued convincingly that the original intent of Article 19(2) contemplates private as well as public interferences.[112] However, this is not, in practical terms, how the global jurisprudence evolved in the decades since the *ICCPR* was first promulgated.

State governments are the primary duty bearers for safeguarding rights.[113] While it is broadly recognized that human rights violations can be committed by private sector actors, such abuses are ultimately understood as a governmental failure.[114] If an oil company, for example, employs thugs to attack environmental campaigners, the abuse may have been perpetrated by the company, but it is the government which failed in its responsibility to protect these activists. Likewise, the government will assume a subsequent duty to pursue an appropriate remedy, by investigating and prosecuting the perpetrators. This is consistent with the government's monopoly on coercive powers, such as the right to imprison. This does not, of course, absolve private actors for their own culpability when they are complicit in or responsible for human rights violations.[115] However, the dynamic of

---

[111] U.N. Human Rights Comm., *General Comment 34*, *supra* note 4, ¶ 32.

[112] Molly Land, *Toward an International Law of the Internet*, 54 HARV. INT'L L.J. 393, 446 (2013).

[113] James W. Nickel, *How Human Rights Generate Duties to Protect and Provide*, 15 HUM. RTS. Q. 77 (1993).

[114] *See, e.g.*, Amnesty Int'l, *Honduras/Guatemala: Attacks on the rise in world's deadliest countries for environmental activists* (Aug. 31, 2016), https://www.amnesty.ca/news/hondurasguatemala-attacks-rise-world's-deadliest-countries-environmental-activists.

[115] *Report of the Special Representative of the Secretary General on the issue of human rights and transnational corporations and other business enterprises, John Ruggie: Guiding Principles on Business*



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

responsibility is significantly different, including, because private sector entities operate with more limited powers, and are subject to regulation by governments.

The duty of governments to regulate in this space is further reinforced by the fact that human rights rules often include a specific obligation for States to take action to prevent rights violations by third parties.[116] While this obligation is most commonly applied in the context of physical attacks against human rights campaigners, it also includes components that are rooted in a responsibility to pass laws which facilitate a robust expressive environment, particularly in the realm of broadcast regulation:

> [T]he State may be required to put in place positive measures to ensure that its own actions contribute to the free flow of information and ideas in society, what may be termed "direct" positive measures. This might involve, for example, putting in place a system for licensing broadcasters which helps ensure diversity and limit media concentration. Perhaps the most significant example of this is the relatively recent recognition of the obligation of States to put in place a legal framework to provide for access to information held by public bodies. [references omitted][117]

Despite their power and reach, platforms play a very different role from governments, which demands a different understanding of how human rights values apply to their operations. It would be a huge conceptual leap to apply a governmental model of freedom of expression obligations to private sector platforms. If YouTube decided to transition to a subscription model, such that their videos were only visible to paying customers, this would certainly run counter to the idea of promoting "the free flow of information and ideas." It would also undoubtedly be within their rights as a company to pursue that revenue stream.

In terms of moderating content, platforms, like all private sector actors, also enjoy a robust "right not to speak."[118] Their freedom to curate their medium as they see fit, and to avoid hosting content they dislike, is a well-established aspect of their

*and Human Rights: Implementing the United Nations "Protect, Respect and Remedy" Framework*, UNHRC, 17th Sess, UN Doc A/HRC/17/31 (2011), https://www.ohchr.org/Documents/Issues/Business/A-HRC-17-31_AEV.pdf.

[116] *See, e.g.*, International Covenant on Civil and Political Rights, *supra* note 91, at art. 20(2).

[117] MENDEL, *Restricting*, *supra* note 105.

[118] Ediciones Tiempo S.A. v. Spain, App. No. 13010/87, 62 Eur. Comm'n H.R. Dec. & Rep. 247, 254 (1989); *Advisory Opinion to the Government of Costa Rica* (1986) Inter-Am Ct HR (Ser. A) No 7.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

own freedom of expression rights.[119] Of course, this right is not absolute.[120] However, it creates a substantially different dynamic than one might expect in a governmental context. This gap, between the theoretical framework in which international freedom of expression principles have traditionally been applied, and the practical realities of private sector content moderation, is the focus of the next section.

## III. SQUARING THE CIRCLE

### *A. The Utility of Human Rights Standards*

Despite the baseline contextual challenges in their application to a purely private sector context, international human rights standards are a promising solution to the present crisis of legitimacy impacting platforms' content moderation structures.[121] International freedom of expression standards are virtually the only conceptual framework for assessing the boundaries of acceptable speech which transcends national law.[122] By providing a spectrum of potential responses, as opposed to a single right answer to moderation challenges, human rights standards help to head off a trend towards consolidation or homogeneity across platforms.[123] This is important to preserving the diversity of content enforcement, which is a key feature of the current system.[124] The idea that content which is not fit for Facebook might still find a home on Twitter or Reddit is important insofar as it prevents a single adverse content decision from bearing the weight of a global ban. From this perspective, a conceptual framework which preserves some equivalent to the "margin of appreciation" doctrine, as applied by the European Court of Human Rights to grant leeway to governments in finding solutions which suit local needs, is

[119] United States Telecom Ass'n v. FCC, 855 F.3d 381, 433–34 (D.C. Cir. 2017) (per curiam) (denying en banc review of a decision upholding the FCC's net-neutrality rules) (Kavanaugh, J., dissenting) (arguing that net neutrality rules violated ISPs' First Amendment rights).

[120] For example, governments frequently impose "must-carry requirements" on broadcasters, whereby operators are required to carry particular stations or types of content. These are well-established aspects of a robust broadcasting regulatory framework; *International Mechanisms for Promoting Freedom of Expression, Joint Declaration on Diversity in Broadcasting*, ORG. OF AM. STATES (2007), www.oas.org/en/iachr/expression/showarticle.asp?artID=719&lID=1.

[121] *See* Jonathan Zittrain, *Three Eras of Digital Governance*, SSRN (Sept. 23, 2019), https://ssrn.com/abstract=3458435.

[122] Frans Viljoen, *Exploring the Theory and Practice of the Relationship between International Human Rights Law and Domestic Actors*, 22 LEIDEN J. INT'L L. 177, 179 (2009).

[123] Llansó, *supra* note 62.

[124] Balkin, *supra* note 78.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

vital.[125] The value of a human rights-based framework is not to tell us who is "right," but rather to ensure that each policy is clear, consistently applied, non-discriminatory, and appropriately proportional in their enforcement.

For example, each platform currently has its own definition of what counts as "hate speech." For Facebook, this is defined as "a direct attack on people based on what we call protected characteristics—race, ethnicity, national origin, religious affiliation, sexual orientation, caste, sex, gender, gender identity, and serious disease or disability" and with "some protections for immigration status."[126] Twitter bans "hateful conduct," rather than hate speech specifically, a category which is defined inclusively as incorporating violent threats, calling for harm against a person or groups, inciting fear, using slurs, or referencing violent events.[127] Google provides no working definition at all, beyond "promoting violence or hatred against individuals or groups" on the basis of a list of specific attributes.[128]

While there is value to this diversity, platforms and their users would greatly benefit from building on the decades of established jurisprudence that human rights law provides, as an avenue to making their respective enforcement systems more resilient, reliable, and transparent than the current paradigm, which is heavily weighted towards speed and efficiency. However, in order to properly implement international freedom of expression concepts, these principles must first be adapted to the specific context in which platforms operate. The natural starting point would be the three-part test for legitimate restrictions on speech, as spelled out in the ICCPR. This requires that any restrictions must be provided by law, imposed for a legitimate purpose, and consistent with fundamental principles of necessity and proportionality. Each aspect of this test requires some reconsideration to apply to the context of platforms.

### *B. Provided by Law*

The first branch of the test, that restrictions be provided by law, imposes standards of clarity, accessibility, transparency, and predictability on both the

---

[125] Hertel v. Switzerland, 59 Eur. Ct. H.R. (ser. B) (1998), § 46; *see also* Tolstoy-Miloslavsky v. United Kingdom, (1995) 20 EHRR 442; markt intern Verlag GmbH and Klaus Beerman v. Germany, 3 Eur. Ct. H.R. (ser. A) (1989).

[126] *Hate Speech*, FACEBOOK (2019), https://www.facebook.com/communitystandards/hate_speech.

[127] *The Twitter Rules*, *supra* note 51.

[128] *Hate Speech Policy*, YOUTUBE HELP, https://support.google.com/youtube/answer/2801939?hl=en.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

dividing line between acceptable and unacceptable speech, and in the application and enforcement of these rules.[129]

In some ways, this is a relatively easy standard to apply to a private sector context, since the platforms' content moderation structure engages many of the same values. It is equally important in the context of "platform law" for individuals to understand the boundaries of acceptable speech, so that they may regulate their conduct appropriately. If it is not clear what the rules are, even enforcement actions that are carefully thought out will appear capricious or arbitrary, which hurts the perceived legitimacy of the system.

One obvious starting point is for platforms to narrow and clarify their content rules. Rather than relatively open-ended statements, such as a commitment to "take action" against "excessively aggressive insults that target an individual, including content that contains slurs or similar language" but "not action against every instance where insulting terms are used,"[130] the content standard could be amended to include a clear definition of prohibited speech, and a specific description of the range of appropriate responses depending on the perceived severity of the offence. The purpose would be to minimize discretion and unpredictability, and allow all stakeholders, including end-users, governments, moderators, and advertisers, to get a clear and comprehensive picture of the boundaries of acceptable speech based on the language. The global nature of platforms means that all of this information should be translated as widely as possible, to facilitate their accessibility among smaller linguistic communities.[131]

Clarity and accessibility are equally important in understanding how the rules are applied, including the procedures for flagging particular content, or appealing against an adverse decision. More broadly, it is important to shed light on the entire range of moderation tools that platforms have at their disposal. This means not just explaining how a platform decides whether to delete particular content, but also determinations to promote or suppress it, and the range of subtler tools that platforms employ to curate information flows.

---

[129] U.N. Human Rights Comm., *General Comment 34*, *supra* note 4, ¶ 25.

[130] *The Twitter Rules*, *supra* note 51.

[131] According to a Reuters report from April 2019, Facebook's content standards are available in 41 languages, though the platform itself is available in 111 languages. Maggie Fick & Paresh Dave, *Facebook's Flood of Languages Leave it Struggling to Monitor Content*, REUTERS (Apr. 23, 2019), https://www.reuters.com/article/us-facebook-languages-insight/facebooks-flood-of-languages-leave-it-struggling-to-monitor-content-idUSKCN1RZ0DW.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

Civil society have long called for more information into how content moderation actually works.[132] This is an obvious advocacy focus, as greater understanding of how the systems operate is a precondition to presenting meaningful recommendations for their improvement. To the platforms' credit, transparency reporting with regard to complaints and enforcement has now become commonplace.[133] However, thus far the platforms' moves toward releasing more data has been seen by many as insufficient.[134] Academic and civil society researchers have demonstrated significant gaps in disclosure in these reports, as well as inconsistencies in how different types of takedowns are characterized, and a continuing lack of clarity regarding enforcement.[135]

From a human rights perspective, it should not be surprising that the current reporting mechanisms have proven ineffective at fostering public trust. It would clearly be insufficient if public accountability over, say, the police force, was limited to analyzing their press releases and published statistics. Robust public oversight, where it exists, takes place not merely through periodic decisions by the agencies to release data, but also through a public right to request information, which transforms the oversight process into an interactive dialogue, rather than a one-sided process of disclosure.[136]

The right to request information manifests in different ways across different legal contexts.[137] However, over the past two decades, it has become firmly entrenched in global understandings of freedom of expression as a human right.[138]

---

[132] *See, e.g.*, *Santa Clara Principles on Transparency and Accountability in Content Moderation* (May 7, 2018), https://santaclaraprinciples.org; *see also* DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET (Colum. Glob. Rep. 2019); *see also* European Commission, *Tackling Illegal Content Online* (Sept. 28, 2017), https://ec.europa.eu/ digital-single-market/en/news/ communication-tackling-illegal-content-online-towards-enhanced-responsibility-online-platforms.

[133] *See, e.g.*, *Facebook's Community Standards Enforcement Report*, FACEBOOK, https:// transparency.facebook.com/community-standards-enforcement.

[134] Sergei Hovyadinov, *Toward a More Meaningful Transparency: Examining Twitter, Google, and Facebook's Transparency Reporting and Removal Practices in Russia*, SSRN (Nov. 30, 2019), https://ssrn.com/abstract=3535671; Nathalie Maréchal & Ellery Roberts Biddle, *It's Not Just the Content, It's the Business Model: Democracy's Online Speech Challenge*, NEW AM: RANKING DIG. RIGHTS (Mar. 17, 2020), https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/.

[135] *Id.*

[136] TOBY MENDEL, FREEDOM OF INFORMATION: A COMPARATIVE LEGAL SURVEY (2003) [hereinafter MENDEL, FREEDOM].

[137] In the United States, this right is generally managed under the Freedom of Information Act, 5 U.S.C. § 552 (1967).

[138] *See, e.g.*, Claude Reyes and Others v. Chile, (2006) Inter-Am. Ct. HR (Ser. C) No. 151; Társaság A Szabadságjogokért v. Hungary, No. 37374/05, [2009] ECHR 618, 53 EHRR 3.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

In its essence, the right to information grants stakeholders an ability to file requests with public bodies for the release of any material under their control, subject to limited and specific exceptions to disclosure which protect core public interests.[139]

In some countries, national right to information legislation already obligates private companies to respond to information requests in certain circumstances.[140] Likewise, the Internet Corporation for Assigned Names and Numbers (ICANN), a California-based, not-for-profit public-benefit corporation which coordinates the global domain name system, has a Documentary Information Disclosure Policy built into its bylaws, which is essentially based on a governmental model of facilitating public requests.[141] Many international financial institutions, such as the World Bank and the Inter-American Development Bank, have their own mechanisms for facilitating access to information requests.[142] It may seem ambitious to envision a parallel commitment from platforms, that information related to content moderation would be made open by default and publicly available upon request. But such bold thinking is necessary in order to foster a level of trust and legitimacy concomitant to their enormous power and influence over the public discourse.

None of this is to suggest that content moderation should take place in an environment of total transparency. As commercial entities, platforms will have a legitimate need to protect their own commercially sensitive information or trade secrets. Likewise, it is possible that the release of certain information about the moderation process could undermine the efficacy of enforcement systems, or enable users to game the system to promote their own content. There will also be privacy

---

[139] *See* MENDEL, FREEDOM, *supra* note 136.

[140] S. AFR. CONST., Promotion of Access to Information Act, 2000. South Africa's law, which applies relatively broadly to the private sector, is something of an outlier, but it is relatively common for such laws to apply to private companies which accept public funding (to the extent of that funding), or which perform a public function. For example, Mexico's General Act of Transparency and Access to Public Information applies to "any individual, legal entity or union who receives and uses public resources or performs acts of authority of the Federation, the States and the municipalities." General Act of Transparency and Access to Public Information, Código Civil, art. 23, Diario Official de la Federación [DOF] 20-04-2015 (Mex.).

[141] ICANN, *ICANN Documentary Information Disclosure Policy* (Feb. 2, 2012), https://www.icann.org/resources/pages/didp-2012-02-25-en.

[142] W. BANK, *Bank Policy: Access to Information* (July 1, 2015), pubdocs.worldbank.org/en/393051435850102801/World-Bank-Policy-on-Access-to-Information-V2.pdf; INTER-AMERICAN DEV. BANK, *Access to Information Policy* (Apr. 26, 2010), idbdocs.iadb.org/wsdocs/getdocument.aspx?docnum=35167427. For a broader discussion of transparency policies at international institutions; *see* Michael Karanicolas, *Openness Policies of the International Financial Institutions: Failing to Make the Grade with Exceptions*, CENTRE L. & DEMOC. (Jan. 10, 2012), https://www.law-democracy.org/wp-content/uploads/2012/01/IFI-Research-Online-HQ.pdf.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

concerns against disclosing certain information related to individual moderation decisions.

However, these same challenges also exist in the context of governments, international financial institutions, and ICANN. Right-to-information laws contain exceptions to disclosure that address these issues. For example, they commonly contain an exception against disclosure of material which would cause commercial harm, either to the government itself or to outside institutions which, in confidence, supplied the government with sensitive information.[143] Another common exception to disclosure is for information which might be expected to undermine the efficacy of law enforcement or regulatory enforcement.[144] Likewise, right-to-information legislation commonly includes an exception to disclosure for material whose release would impact personal privacy.[145] While these rules are not always an exact match for the specific context that platforms operate in, they could be adapted to purpose quite easily. Governments, quasi-public entities and international financial institutions, around the world are able to maintain the integrity of their activities while adhering to principles that information should be considered open by default and available upon request, subject to limited and specific exceptions. There is no reason why platforms could not adhere to parallel standards, as far as information connected to their content moderation processes are concerned.

### *C. Serving a Legitimate Purpose*

The second branch of the three-part test, that limitations on speech must serve a legitimate purpose, is more difficult to translate, since it restricts the grounds under which governments may impose restrictions. Here platforms would need to be accorded much wider discretion, given their inherent right to promote discourse of a particular character. If, for example, a platform was created with the explicit purpose of facilitating sports-related discussions, they might want to impose a rule that any off-topic conversations (those not related to sports) would be subject to deletion. Such a broad rule, while obviously inappropriate for governments to impose on the general population, would be fine in a more limited private sector context. More

[143] For example, Romania exempts from disclosure "information about financial or commercial activities if, according to the law, its release is detrimental to the principle of fair competition." Law 544/12 on Free Access to Public Information (2001) (Rom.). For the American approach to the intersection of commercially sensitive matters with the Freedom of Information Act, see John Delaney, *Safeguarding Washington's Trade Secrets: Protecting Businesses from Public Records Requests*, 92 WASH. L. REV. 1905 (2017).

[144] *See, e.g.*, Access to Information Act, R.S.C. 1985, c. A-1, s. 16(1) (Can.).

[145] *See id.* at 19(1).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

generally, platforms face commercial pressures in regulating speech which distinguish them from governments.

However, there is still value in applying this branch of the test, insofar as it mandates that restrictions should be established and applied "in the light of universality of human rights and the principle of non-discrimination."[146] In practical terms, this might be understood as imposing standards of consistency and coherence, supporting a need for platforms to apply their rules in a manner which is consistent with that purpose, and which does not improperly discriminate, for example by judging LGBTQ expressions of intimacy or sexuality against a harsher standard than parallel heterosexual expressions.

More generally, the second branch of the test requires that restrictions on speech be connected to some sort of distinct and overarching policy rationale. In the case of a government, this might mean justifying a restriction on tobacco advertising by connecting it to the broader aim of protecting public health. In the case of platforms, while the categories of restriction may be more open-ended, the requirement to cite some underlying purpose for the restriction is nonetheless valuable to moving the enforcement system away from appearing discretionary and capricious. It is also a necessary precondition for assessing the impact of the rule against its benefits, in line with the third branch of the test.

### *D. Necessary and Proportionate*

The third branch of the three-part test, which focuses on the necessity of the restriction to achieve the objective and the proportionality of the impact on freedom of expression as set against this goal, is where much of the heavy lifting is traditionally done in assessing potential violations of the ICCPR, or parallel regional treaties such as the European Convention on Human Rights. Inherent in this branch of the test is an understanding that government limits on speech should be tailored as narrowly as possible in order to achieve their objective.[147] In other words, there is an implicit assumption that a light-touch approach, with minimal interferences in the public discourse, is preferable.

The problem with applying this framework to the context of major platforms is that, in practical terms, there are technical challenges which effectively preclude a hands-off approach to moderation. While some platforms may appear to be less active than others in deleting controversial content, this does not mean that moderation is not taking place. At a bare minimum, every platform acts to combat

[146] U.N. Human Rights Comm., *General Comment 34*, *supra* note 4, ¶ 32.

[147] *Id.* ¶¶ 33–34.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

known threats such as the distribution of malware, or to remove spam, as part of their routine security posture.[148] These efforts are necessary to keep the services usable on a basic level. Likewise, when YouTube selects which videos to auto-play for users, or which videos appear on its homepage, or when Facebook curates which posts appear in a person's News Feed, this is a form of moderation. While these decisions may be subtler than deletions, they can be incredibly consequential to the spread or suppression of particular messages. There are a few platforms which do not carry out this kind of moderation, but these tend to be relatively fringe exceptions.[149] For global platforms, the vast firehose of content which they host requires a relatively strong hand in managing how information is presented to users.

In this sense, content moderation may be considered as more comparable to a positive freedom of expression obligation, rather than a negative restriction on speech.[150] A good analogy to understanding this is how international human rights law treats broadcast regulation. In contrast to the light-touch approach favored for most forms of media, international human rights law has long recognized that a stronger regulatory hand is necessary in the broadcasting space.[151] There are a number of rationales for this distinction, including the notion that spectrum is a limited resource which must be rationed, and that government intervention is necessary in order to prevent broadcasters from interfering with one another's transmissions.[152] However, in essence, the principle is that the nature of the broadcast medium is such that government intervention in the space is necessary for it to work. This pulls the calculus away from a traditional necessity and proportionality test, whereby freedom of expression is balanced against a competing value (such as

---

[148] *See, e.g.*, Shabnam Shaik, *Improvements in protecting the integrity of activity on Facebook*, FACEBOOK (Apr. 12, 2017), https://www.facebook.com/notes/facebook-security/improvements-in-protecting-the-integrity-of-activity-on-facebook/10154323366590766; *see also Malware Checkpoint for Facebook*, FACEBOOK (July 10, 2012), https://www.facebook.com/notes/facebook-security/malware-checkpoint-for-facebook/10150902333195766/.

[149] *See, e.g.*, *AP Explains: What is the online forum 8chan?*, ASSOCIATED PRESS (Aug. 5, 2019), https://apnews.com/39b7182524204189bbffab516a9df393; *see also* Sarah Emerson, *Founder of Voat, the 'Censorship-Free' Reddit, Begs Users to Stop Making Death Threats*, VICE (Apr. 26, 2019), https://www.vice.com/en_us/article/gy4gzy/founder-of-voat-the-censorship-free-reddit-begs-users-to-stop-making-death-threats.

[150] MENDEL, *Restricting*, *supra* note 105.

[151] *See, e.g.*, *Access to the Airwaves Principles on Freedom of Expression and Broadcast Regulation*, ART. 19 (Mar. 2002), https://www.article19.org/data/files/pdfs/standards/accessairwaves.pdf; *see also* Recommendation 2007(2) of the Committee of Ministers of the Council of Europe on Media Pluralism and Diversity of Media Content (Jan. 31, 2007), https://search.coe.int/cm/Pages/result_details.aspx?ObjectId=09000016805d6be3.

[152] *See, e.g.*, Red Lion Broadcasting Co. v. FCC, 395 U.S. 367 (1969).



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

national security),[153] and towards an analysis of the competing expressive costs and benefits of a particular intervention, to ensure that the resulting framework is beneficial to freedom of expression.[154]

As applied to the context of platforms, this means that moderation decisions should not be assessed as an "interference" with freedom of expression, any more than a decision to grant a broadcasting license to one applicant over another is assessed an interference with the latter's rights. In that case, the general understanding that governments have an obligation to allocate spectrum, which in turn requires denying space to some applicants, would mean that a human rights analysis would focus on whether the system to allocate licenses as a whole served the public interest, and whether it was administered and enforced in a fair, accountable and transparent manner. Rather than considering whether a moderation policy or decision is strictly necessary to achieve a given aim, applying the necessary and proportionate standard in the context of content moderation at platforms requires an assessment of whether the structure adequately respects, protects and promotes the expressive interests of all of the parties involved, as well as the broader public. This means a careful consideration of the impacts of a particular moderation or enforcement posture on freedom of expression, both for good and for ill, in order to maximize the promotional and positive aspects while minimizing harmful or restrictive ones.

For example, in recent years there has been an increasing move to automate content moderation structures, including filtering technologies which flag and delete material at the point of upload.[155] In the context of governmental restrictions, automated content filtering systems are generally viewed as an unjustifiable interference with freedom of expression.[156] The rationale for this is that filtering systems, by proactively removing content at the point of creation, are effectively analogous to a form of prior censorship, which itself is generally unacceptable according to international human rights standards.[157]

---

[153] Alec Stone Sweet & Judkins Mathews, *Proportionality Balancing and Global Constitutionalism*, 47 COLUM. J. TRANSNAT'L L. 68 (2008).

[154] For an application of this calculus to the field of copyright law, another area where restrictions on speech are generally employed for the sake of promoting speech, see Michael Karanicolas, *Reconceptualising Copyright: Adapting the Rules to Respect Freedom of Expression in the Digital Age*, CENTRE L. & DEMOC. (2013), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3423085.

[155] *See, e.g.*, Leron Solomon, *Fair Users or Content Abusers? The Automatic Flagging of Non-Infringing Videos by Content ID on YouTube*, 44 HOFSTRA L. REV. 237 (2015) (discussing the persistent failure to apply fair use doctrine in alleged cases of copyright infringement).

[156] *Joint Declaration on Freedom of Expression and "Fake News", Disinformation and Propaganda*, *supra* note 65, at ¶ 1.g.

[157] *Joint Declaration on Freedom of Expression and the Internet*, ORG. FOR SEC. & CO-OPERATION IN EUR. 68 (June 1, 2011), www.osce.org/fom/99558?download=true.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

But in the context of platforms, and of understanding moderation structures as necessary to support freedom of expression, a proportionality assessment presents a more nuanced picture. Given the scale at which platforms operate, the speed with which information spreads, and the impact that abandoning proactive assessments might have on, say, the ability to detect malware and spam, there is an argument to be made that abandoning these filters would actually be broadly harmful to freedom of expression, since it would severely degrade the expressive environment.

However, when one considers the broader impacts of automated filtering for content which is more contextual, such as hate speech or "terrorist speech," the calculus becomes more difficult. One could still make an argument that letting such material run rampant would be counter to freedom of expression, since users will not want to engage in a platform which is overridden with toxicity. However, in considering the broader impacts of such a system, one would also have to consider that it is vastly more difficult to automatically identify hate speech or "terrorist speech," or for that matter abuse or harassment, than it is to identify malware or child sexual abuse material. This leads to a greater potential for negative impacts, since there will necessarily be more collateral damage, or "by-catch," to such a system. As a consequence, one might conclude that applying automated systems to these categories of content would require greater safeguards in order to satisfy this branch of the test. This might include limiting automated takedowns to content which has been pre-identified as prohibited (as opposed to allowing automated systems to independently assess whether material is prohibited), or requiring human confirmation for takedowns, or some form of random auditing by a human.

It is worth noting that this calculus is not static, and can shift over time or in response to emerging events. For example, there have been a number of recent incidences where platforms have imposed emergency moderation measures in response to a pressing crisis.[158] Where such emergency measures have been imposed, it is reasonable to carry out an assessment within the specific emergency context that exists, though it is equally imperative to reassess these factors once the situation has passed, rather than allowing them to become the default way of doing things.

To be clear, the preceding arguments are not intended to make a broader case for assessing government interferences into digital speech as analogous to broadcast regulation, with all the leeway that that implies. Nor do they resolve human rights challenges related to the indirect nature of governmental enforcement actions that are currently being "laundered" through private sector platforms. However, in developing an assessment mechanism for decisions which are taken independently

[158] *See, e.g.*, Vijaya Gadde & Matt Derella, *An update on our continuity strategy during COVID-19*, TWITTER (Mar. 16, 2020), https://blog.twitter.com/en_us/topics/company/2020/An-update-on-our-continuity-strategy-during-COVID-19.



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

by platforms, this modified version of the three-part test helps to square the circle between human rights law and platform law, and allows platforms to apply a robust and extensive volume of international standard-setting work in this space to better inform and guide their decisionmaking.

## CONCLUSION

This Article is not intended to provide a comprehensive definition of what policies platforms should adopt to fulfil their obligations as global gatekeepers for online speech. In considering each policy at each platform, it will be necessary to undertake a careful and contextual analysis, which will in turn require an in-depth understanding of the expressive environment which the platform itself hosts, and the impact that every decision has on curating that environment. At the moment, only the platforms themselves have access to the depth of information which might be necessary to carry out such an analysis. Hopefully, as conversations around improving global content moderation structures move forward, and academics and civil society experts continue to engage on these issues, that dynamic will change, as more global expertise is brought to bear on resolving the thorny problems of defining appropriate boundaries for online speech in the platform realm.

The internet has, since its early days, been viewed as a great disruptor, unleashing enormous changes on a range of industries, and transforming the lives of vast numbers of people. The nature of online speech has been at the forefront of these impacts, driving a concomitant need to radically rethink the rules around freedom of expression. But while regulatory systems may need to be revised, the core values that underlie them continue to reflect a carefully constructed balance, which has developed through the engagement of many of the smartest minds of the past century. These principles have weathered a number of storms and challenges before, and emerged more resilient and vital to democratic culture than ever.

This Article does not suggest turning back the clock. It is clear that platforms' own content decisions are now an enormously influential component of the global public discourse. Moreover, there are important benefits to this, such as the enduring ability of people living in repressive corners of the world to use online technologies as a means of skirting the repressive restrictions that surround traditional communication. However, this thrusts enormous responsibilities onto the shoulders of private sector platforms, who now bear a level of influence that rivals, and in some cases even exceeds, that of nation States. While platforms are still conceptually different from governments, it is clear that a rights-based framework presents the clearest and most effective avenue for supporting freedom of expression in the application of their policies. While this Article does not portend to provide a complete definition for how platforms can manage their role in accordance with human rights principles, it aims to provide an initial mapping of the considerations



Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu

that should support this conversation going forward, and guide the application of international freedom of expression standards to the context of platform law.





Volume XX – 2019-2020 ● ISSN 2164-800X (online)
DOI 10.5195/tlp.2020.236 ● http://tlp.law.pitt.edu