VOLUME 62, NUMBER 1, WINTER 2021

Governing the Global Public Square

Rebecca J. Hamilton*

Social media platforms are the public square of our era—a reality that has been entrenched by the widespread closure of physical public spaces in response to the COVID-19 pandemic. This online space is global in nature, with over 3.6 billion users worldwide, but its governance does not fall solely to governments. With the rise of social media, important decisions about what content does—and does not—stay online are made by private technology companies. Reflecting this reality, cutting-edge scholarship has converged on a triadic approach to understanding how the global public square operates—with states, users, and technology companies marking out three points on a "free speech triangle" that determines what content appears online. While offering valuable insights into the nature of online speech regulation, this scholarship, which has influenced public discussion, has been limited by drawing primarily on a recurring set of case studies arising from the United States and the European Union. As a result, the free speech triangle has locked in assumptions that make sense for the United States and the E.U., but that regrettably lack broad applicability.

This Article focuses our attention on the global public square that actually exists, rather than the narrow U.S.- and European-centric description that has commanded public attention. Drawing on interviews with civil society, public sources, and technology company transparency data, it introduces a new set of case studies from the Global South, which elucidate important dynamics that are sidelined in the current content moderation discussion.

Drawing on this broader set of materials, I supplement the free speech triangle's analysis of who *is responsible for online content, with the question of* what *these actors do. In this way, activity within the global public square can be grouped into four categories: content production, content amplification, rule creation, and enforcement. Analyzing the governance of the global public square through this functional approach preserves important insights from the existing literature while also creating space to incorporate the plurality of regulatory arrangements around the world. I close with prescriptive insights that this functional approach offers to policymakers in a period of unprecedented frustration with how the global public square is governed.*

INTRODUCTION

More than 3.6 billion people across the globe regularly spend time on social media.[1] Social media platforms are the new public square—and there is widespread discontent about what speech does and does not get to flow

[*] Associate Professor of Law, American University, Washington College of Law. Heartfelt thanks to Janea Marks and Angela Gerrits for research assistance, and to the HILJ editors for their dedication to the editing process, notwithstanding the impact of COVID-19, and for their deft work to help me incorporate the new developments that kept unfolding daily as this Article was set to go to print. I am grateful to Hannah Bloch-Wehba, Gabriella Blum, Pamela Bookman, Jay Butler, Julie Cohen, Danielle Citron, Jennifer Daskal, Evelyn Douek, Matthew Erie, Claudia Haupt, Andrew Ferguson, Susan Franck, Maggie Gardner, Rebecca Ingber, Daphne Keller, Molly Land, Martha Minow, David Pozen, Barrie Sander, and David Sloss for comments on earlier versions. All errors are my own.

[1] *Number of Social Media Users Worldwide,* STATISTA, https://www.statista.com/statistics/278414/number-of-worldwide-social-network-users/ [https://perma.cc/4D7B-UAU2] (last visited July 15, 2020).

freely within it.[2] In the United States, members of Congress have questioned the role of Facebook in facilitating Russian interference in the 2016 presidential election.[3] President Trump accused various platforms of political bias, claiming that they interfered with his re-election campaign, and engaged in voter suppression.[4] In the wake of recent protests for racial justice, activists have started to call for new "rules of engagement" around sharing the videos and images of the deaths of Black Americans.[5]

Looking abroad, in Syria, lawyers have been frustrated that YouTube has removed videos that could serve as evidence in war crimes prosecutions.[6] In Brussels, lawmakers have chastised Twitter for being too slow to remove hate speech.[7] In Myanmar, human rights leaders have asked why Facebook

[2] As Danielle Citron and Neil Richards argued so convincingly in a 2018 article, social media platforms do not function as an open democratic forum in the way that U.S. audiences, at least, have traditionally thought of a "public square." *See* Danielle K. Citron & Neil M. Richards, *Four Principles for Digital Expression*, 95 WASH. U. L. REV. 1353, 1356 (2018) ("[Private companies] block, filter, mute, and decrease the visibility of online expression, making it difficult for some to engage in public discourse."). And, of course, they are part of an internet ecosystem that is also so much more than a mere public square. *Id.* at 1355.

[3] *Facebook Boss Mark Zuckerberg Tells Congress the Company is in an "Arms Race" with Russia*, ABC NEWS (Apr. 11, 2018, 4:21 PM), https://www.abc.net.au/news/2018-04-11/facebook-boss-mark-zuckerberg-fronts-united-states-congress/9639764 [https://perma.cc/6DJH-YSS8].

[4] *See, e.g.*, Exec. Order No. 13,925, 85 Fed. Reg. 34,079 (June 2, 2020); President Donald J. Trump, FACEBOOK (Nov. 6, 2020, 5:51) ("Twitter is out of control, made possible through the government gift of Section 230!"); Davey Alba et al., *Twitter Adds Warnings to Trump and White House Tweets, Fueling Tensions*, N.Y. TIMES (June 3, 2020), https://www.nytimes.com/2020/05/29/technology/trump-twitter-minneapolis-george-floyd.html [https://perma.cc/53J8-M5PR]; Cat Zakrzewski, *The Technology 202: Trump Escalates His War with Twitter . . . on Twitter*, WASH. POST (May 27, 2020), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/05/27/the-technology-202-trump-escalates-his-war-with-twitter-on-twitter/5ecd56e7602ff165d3e417fc [https://perma.cc/65TY-M8WA]; Tony Romm & Elizabeth Dwoskin, *Trump Signs Order That Could Punish Social Media Companies for How They Police Content, Drawing Criticism and Doubts of Legality*, WASH. POST (May 28, 2020, 9:48 AM), https://www.washingtonpost.com/technology/2020/05/28/trump-social-media-executive-order/ [https://perma.cc/U8GB-L87G]; Cat Zakrzewski, *The Technology 202: Snap's Decision Could Have Implications for Trump's Young Voter Outreach*, WASH. POST (June 4, 2020, 9:05 AM), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/04/the-technology-202-snap-s-decision-could-have-implications-for-trump-s-young-voter-outreach/5ed7dd6288e0fa32f82315d1 [https://perma.cc/W5MB-UB56].

[5] Melanye Price, Opinion, *Please Stop Showing the Video of George Floyd's Death*, N.Y. TIMES (June 3, 2020), https://www.nytimes.com/2020/06/03/opinion/george-floyd-video-social-media.html [https://perma.cc/3UUP-RFEQ]; Craig Timberg, *Zuckerberg Acknowledges Facebook Erred by Not Removing a Post that Urged Armed Action in Kenosha*, WASH. POST (Aug. 28, 2020, 4:58 PM), https://www.washingtonpost.com/technology/2020/08/28/facebook-kenosha-militia-page/ [https://perma.cc/2TKV-K9M7] (explaining the context of the Facebook CEO's admission in an internal meeting that despite the event garnering many user complaints it remained until after two protesters were shot); *see also* Rachel Lerman, *YouTube Is Starting to Age-Restrict News Videos of George Floyd's Killing*, WASH. POST (June 4, 2020, 4:53 PM), https://www.washingtonpost.com/technology/2020/06/04/youtube-restricts-floyd-videos [https://perma.cc/6VZA-YZBA] (explaining that YouTube recently began displaying age-appropriateness warnings on videos of the killing of George Floyd).

[6] Bernhard Warner, *Tech Companies Are Deleting Evidence of War Crimes,* ATLANTIC (May 8, 2019), https://www.theatlantic.com/ideas/archive/2019/05/facebook-algorithms-are-making-it-harder/588931/ [https://perma.cc/HAW4-3ZUH].

[7] Mark Scott, *Twitter Fails E.U. Standard on Removing Hate Speech Online,* N.Y. TIMES (May 31, 2017), https://www.nytimes.com/2017/05/31/technology/twitter-facebook-google-europe-hate-speech.html [https://perma.cc/W8ML-7SCD].

has allowed speech that incites violence to remain online, enabling atrocities to flourish.[8]

In the face of society-wide "lockdowns" in response to the COVID-19 pandemic, such criticisms of social media platforms gain further salience; the spaces these platforms create and curate currently constitute not only a *new* public square, but de facto the *only* public square in many locales worldwide. Even during the wave of international protests in the physical public square in the wake of the George Floyd murder and reinvigoration of the Black Lives Matter movement, the online public square has continued to serve as a place of organization and activism for those unable to join the physical protests.[9] It is also, however, a place for conspiracy theorists to gain adherents and for false claims to go viral.[10] In response, platforms are developing new content moderation guidelines, such as removing and banning specific conspiracy theories, adding fact-check labels to politicians' posts, disabling accounts that incite violence, and collaborating with and prioritizing posts of various public health agencies.[11]

[8] Letter from Six Civil Society Groups in Myanmar to Mark Zuckerberg, Chief Exec. Officer, Facebook (Apr. 5, 2018) (available at https://drive.google.com/file/d/1Rs02G96Y9w5dpX0Vf1LjWp6B9mp32VY-/view [https://perma.cc/PZJ7-MHKG]); *see also In Myanmar, Facebook struggles with a deluge of disinformation,* THE ECONOMIST, Oct. 22, 2020, https://www.economist.com/asia/2020/10/22/in-myanmar-facebook-struggles-with-a-deluge-of-disinformation [https://perma.cc/KZ2T-MW2P] (noting an increasing volume of hate speech on Facebook in Myanmar in the build up to the country's November 2020 election.).

[9] Michelai Graham, *How Activist Group Freedom Fighters DC Mobilized Using Social Media*, TECHNICAL.LY (June 9, 2020, 6:05 PM), https://technical.ly/dc/2020/06/09/heres-how-activist-group-freedom-fighters-dc-mobilized-using-social-media/ [https://perma.cc/FC3L-CY9T] (describing the new Freedom Fighters DC, who organized a protest in under a week of formation that was attended by over 1,000 people); *#OnlineProtest A 7-Day Nonviolent Livestream*, KING CENTER (June 2, 2020), https://thekingcenter.org/2020/06/02/onlineprotest-for-social-justice-a-7-day-nonviolent-livestream-initiative [https://perma.cc/B9N8-M9Y9]; Alaa Elassar, *This Website Helps People with Illnesses and Disabilities Participate in Black Lives Matter Protests*, CNN (Sept. 27, 2020, 5:45 PM) https://www.cnn.com/2020/09/27/us/online-protest-public-public-address-trnd/index.html [https://perma.cc/276X-2QSK]; Caroline Fox, *14 Ways to Support Black Lives Matter Protests if You Can't Be There in Person*, INSIDER (Jul. 21, 2020, 10:17 AM) ("The Guardian reported that 22,000 people from all over the world tuned in to the online rally using Zoom, Facebook Live, and Instagram.").

[10] *See e.g.,* Elizabeth Dwoskin, *Misinformation About Coronavirus Finds New Avenues on Unexpected Sites*, WASH. POST (May 20, 2020, 6:00 AM), https://www.washingtonpost.com/technology/2020/05/20/misinformation-coronavirus-plandemic-workaround/ [https://perma.cc/QEK2-X84R] (YouTube began implementing rules banning "Plandemic" misinformation and prioritizing CDC and WHO posts, but users uploading links to video began to include links to the CDC website to confuse algorithm).

[11] Davey Alba, *Virus Conspiracists Elevate a New Champion*, N.Y. TIMES (May 9, 2020), https://www.nytimes.com/2020/05/09/technology/plandemic-judy-mikovitz-coronavirus-disinformation.html [https://perma.cc/6LQ5-9P79] (Facebook and YouTube's removal of "plandemic" videos, Twitter's blocking of the pandemic related hashtags #plagueofcorruption #plandemicmovie); Raymond Zhong et al., *Behind China's Twitter Campaign, a Murky Supporting Chorus*, N.Y. TIMES (June 8, 2020), https://www.nytimes.com/2020/06/08/technology/china-twitter-disinformation.html [https://perma.cc/MU9H-59C2] (fact checking China's foreign ministry spokesman on claim that COVID-19 came to China via U.S. military); Craig Timberg et al., *Men Wearing Hawaiian Shirts and Carrying Guns Add a Volatile New Element to Protests*, WASH. POST (June 4, 2020, 12:14 PM), https://www.washingtonpost.com/technology/2020/06/03/white-men-wearings-hawaiian-shirts-carrying-guns-add-volatile-new-element-floyd-protests/ [https://perma.cc/P4JX-UC7W] (Twitter and Facebook report the removal of boogaloo accounts only if the accounts incite violence); Cat Zakrzewski, *The Technology 202: Coronavirus Could Change how*

In the formative years of the internet, legal scholars and pundits alike speculated about what systems would develop to govern a medium that seemed to defy geographically defined organizing principles. Some saw the internet as heralding a new era that would transcend state-based governance.[12] Others contended that states would continue to be the dominant force for establishing and enforcing the rules.[13] At this juncture, those positing a dominant position for states have fared the better of the two, with at least some states playing an outsized role in shaping norms online.[14] But none of these early accounts provide anything close to an accurate picture of how, in practice, online speech is now being regulated. This is because none of the early scholarship foresaw the emergence of social media, and with it,

Social Networks Approach Public Health, WASH. POST (Apr. 20, 2020, 8:57 AM), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/04/20/the-technology-202-coronavirus-could-change-how-social-networks-approach-public-health/5e9cb4ec88e0fa101a7663d6/ [https://perma.cc/G5KF-GHFE] (YouTube began blocking COVID-19 hoaxes and prioritizing CDC and WHO guidelines); Ken Dilanian & Kit Ramgopal*, Facebook Blocks Russia-Backed Accounts, but Other Sites Keep Churning Out Content Aimed At Americans*, NBC News (Oct. 9, 2020, 8:23 AM), https://www.nbcnews.com/politics/2020-election/facebook-blocks-russia-backed-accounts-other-sites-keep-churning-out-n1242683?fbclid=IWAR23qffQBgYT20YRoHYYUk8SqguBGp6SQ9wv4NXkYAFhTtJDZ58TNrXk_GU [https://perma.cc/2CY9-RYL3] (discussing how the company is grappling with Russian-backed media accounts such as Peacedata, which targeted American liberals, and predicting that Russian-led conservative propaganda sites will add to anger and division over claims of voter fraud in the U.S.); David Ingram, *Twitter is Making it More Difficult to View Misleading Information About Voting*, NBC NEWS (Oct. 9, 2020, 12:14 PM) https://www.nbcnews.com/tech/tech-news/twitter-making-it-more-difficult-view-misleading-information-about-voting-n1242720?fbclid=IWAR0MvCEXyZLyWwTOhZGOpmfM6g6zHdIhXuH6hKOHpGu6RV1xgMBNoLNoQnA [https://perma.cc/5G2F-7XUD] (discussing Twitter's plan to place "prominent and obtrusive" warning labels on fact-checked published election-related tweets); David Ingram & Brandy Zadrozny, *"Not Done": Why Social Media's Biggest Election Test is Ahead*, NBC NEWS (Nov. 4, 2020, 7:35 PM) https://www.nbcnews.com/tech/tech-news/not-done-why-social-media-s-biggest-election-test-ahead-n1246506?fbclid=IWAR1VJNne4Q-iiivTjQblmaoCmcezXN8DSTCBV9s8A4vpgWDbGj4kyz4HVhY [https://perma.cc/BUP4-XNFH] (explaining that despite the false information about voting, including misleading videos of a person burning ballots or a poll watcher supposedly being denied entry to a polling station, platforms took actions to use fact-check labeling, ban accounts spreading QAnon conspiracies, disabled specific hashtag searches, and more); Christopher Bing et al., *Spanish-Language Misinformation Dogged Democrats in U.S. Election*, REUTERS (Nov. 7, 2020, 10:38 AM) https://www.reuters.com/article/us-usa-election-disinformation-spanish/spanish-language-misinformation-dogged-democrats-in-u-s-election-idUSKBN27N0ED [https://perma.cc/ENB5-5J2J] (commenting on how the computer matrix and low employee capacity resulted in some slow reactions for platforms, including Facebook, to remove and fact-check Spanish-language posts reflecting the U.S. president's messaging of voter fraud due to mail-in-voting or the QAnon conspiracy theory); Rebecca Klar, *Facebook Cracks Down On Pages Linked to Bannon*, HILL (Nov. 10, 2020, 3:49 PM) https://thehill.com/policy/technology/525365-facebook-cracks-down-on-pages-linked-to-bannon?fbclid=IWAR1YNJb68Lir3IvcPsRSo8otMUbFZ_X8juegvScSHVC1GXNzq-zM02yFr8M [https://perma.cc/4TX7-AZWG].

[12] John Perry Barlow, *A Declaration of the Independence of Cyberspace,* ELECTRONIC FRONTIER FOUND. (Feb. 8, 1996), https://www.eff.org/cyberspace-independence [https://perma.cc/9CGR-ZXSF].

[13] *See* JACK GOLDSMITH & TIM WU, WHO CONTROLS THE INTERNET? ILLUSIONS OF A BORDERLESS WORLD 180–84 (2006).

[14] *See generally* Jennifer Daskal, *Borders and Bits,* 71 VAND. L. REV. 179 (2018) (describing how state regulation of data is having extraterritorial reach); *see also* Editorial, *There May Soon Be Three Internets. America's Won't Necessarily Be the Best*, N.Y. TIMES (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/opinion/internet-google-china-balkanization.html [https://perma.cc/4PLC-K4K3] (describing state-based control over the internet by China, the United States, and the European Union).

the rise of the major technology companies, like Facebook, Twitter, and Google, that run social media platforms.[15]

In the past decade, theoretical efforts have begun to catch up with reality. Leading scholars in the field have moved away from the earlier dyadic state-user models that focused on state regulation to embrace the notion that "Free Speech is a Triangle."[16] A growing consensus has emerged that the determination of what appears online occurs through a process of contestation between three different groups forming the prongs of the free speech triangle—states, users, and technology companies—with the latter newly worthy of academic scrutiny.[17]

While a significant improvement over earlier theories of online speech regulation, this cutting-edge triadic literature faces a serious limitation that this Article seeks to overcome: The literature is derived almost exclusively from a limited set of examples in the United States and the European Union ("E.U.")—with occasional supplementary examples from China.[18] Without a doubt, these examples have yielded important insights. Looking at how platforms responded to German legislation requiring platforms to remove hate speech under threat of financial penalty has strengthened work on important concepts such as "collateral censorship."[19] Looking at how the U.S. government responded to the disclosure of diplomatic cables by Wikileaks has increased our understanding of the role that soft power can play in the relationship between government and technology companies.[20] However, by

[15] Facebook and Twitter are names of both social media platforms and the technology companies that own them. Google owns the YouTube platform.

[16] Jack M. Balkin, *Free Speech is a Triangle*, Colum. L. Rev. 2011, 2012 (2018) ("Free speech is a triangle. The conception of free expression . . . [in the 19th and 20th centuries] concerned whether nation-states and their political subdivisions would regulate the speech of people living within their borders . . . In the early twenty-first century, freedom of expression increasingly depends on a third group of players: a privately owned infrastructure . . . composed of firms that support and govern the digital public sphere that people use to communicate."). *See also* Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech,* 131 Harv. L. Rev. 1599, 1603 (2018) (referencing a "triadic model of speech"); Kristen E. Eichensehr, *Digital Switzerlands*, 167 U. Pa. L. Rev. 665, 703–12 (2019) (using a "triangular framing" to represent the "cyberspace ecosystem").

[17] *Supra* note 16; *see also* Felix T. Wu, *Collateral Censorship and the Limits of Intermediary Immunity*, 87 Notre Dame L. Rev 293 (2011); Yochai Benkler, *A Free Irresponsible Press: Wikileaks and the Battle over the Soul of a Networked Fourth Estate*, 46 Harv. C.R.-C.L. L. Rev. 311 (2011); Derek E. Bambauer, *Orwell's Armchair*, 79 U. Chi. L. Rev. 863 (2012); Jeffrey Rosen, *The Deciders: The Future of Privacy and Free Speech in the Age of Facebook and Google*, 6 Const. L. Rev. 35 (2013); Tim Wu, The Attention Merchants: The Epic Scramble to Get Inside Our Heads (2016); Alan Z. Rozenshtein, *Surveillance Intermediaries*, 70 Stan. L. Rev. 99 (2018); Daskal, *supra* note 14, at 185 (observing that "the role and power of these third-party players were neither anticipated nor accounted for in the more stylized debates of the 1990s."); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State,* 72 SMU L. Rev. 27 (2019); Lina M. Khan & David E. Pozen, *A Skeptical View of Information Fiduciaries,* 113 Harv. L. Rev. 497 (2019) ("Figuring out how to regulate digital firms such as Facebook, Google, and Twitter is one of the central challenges of the Second Gilded Age.").

[18] *See infra* Part I.A.

[19] *See infra* Part I.A; Jack M. Balkin, *Old-School/New-School Speech Regulation,* 127 Harv. L. Rev. 2296, 2298 (2014) (Balkin describes "collateral censorship" as occurring when the state uses another entity to undertake its censorship goals.).

[20] *See* Benkler, *supra* note 17; Bambauer, *supra* note 17, at 891–94.

failing to look at a more diverse range of real-world examples, the free speech triangle has entrenched a set of assumptions that make sense within the limited examples on which the literature has focused, but do not have broad applicability.

The literature typically assumes, for example, that content moderation decisions made by technology companies are playing out against an offline environment where the rule of law is in operation. Globally speaking, that assumption does not hold.[21] Another assumption within the triadic literature is that the state is a regulator of online content. This is unremarkable if one focuses on, say, activity in Brussels, but when you cast your eye to South Sudan it becomes clear that the state is not always able to take on that role. Some states simply do not have the capacity. Likewise, the assumption that the state's main activity is as a regulator of social media has meant that until very recently this literature has paid less attention to the possibility that the state is primarily a user and *abuser* of social media. This insight seems obvious from the vantage point of 2020, in the aftermath of Robert Mueller's report on Russian inference in the 2016 U.S. presidential election and as states push out misinformation on COVID-19 through social media platforms.[22] Yet the possibility of a state using social media in a concerted and systematic way to achieve a harmful goal could have been seen as far back as 2013 if more attention had been paid to places outside the United States and the E.U. As a result of these and other assumptions, the triadic literature systematically misses or minimizes key dynamics at play in large parts of the globe.[23]

Digital activists, journalists, and social scientists have begun to explore these missing dynamics.[24] Critical race theorists and computer scientists are

[21] *See infra* Part II.

[22] *See, e.g.*, Robert S. Mueller III, U.S. Justice Dep't, Report On The Investigation Into Russian Interference in the 2016 Presidential Election (2019); Betsy Woodruff Swan, *State Report: Russian, Chinese and Iranian Disinformation Narratives Echo One Another*, Politico (Apr. 21, 2020, 04:30 AM), https://www.politico.com/news/2020/04/21/russia-china-iran-disinformation-coronavirus-state-department-193107 [https://perma.cc/YQQ5-VG5V]; Sheera Frenkel et al., *Surge of Virus Misinformation Stumps Facebook and Twitter*, N.Y. Times (June 1, 2020), https://www.nytimes.com/2020/03/08/technology/coronavirus-misinformation-social-media.html [https://perma.cc/YEK8-6RF9] (explaining that the WHO refers to the viral misinformation from independent users and state actors alike as an "infodemic").

[23] Accounts focusing on users in the United States and Europe miss the majority of social media users globally. *See, e.g.*, Press Release, Facebook, Facebook Reports Second Quarter 2020 Results (June 30, 2020) (available at https://s21.q4cdn.com/399680738/files/doc_financials/2020/q2/Q2'20-FB-Financial-Results-Press-Release.pdf [https://perma.cc/J3EP-SE44]) (showing that in Q4 2018, Facebook had 468 million active daily users in the United States, Canada and Europe, and 1.05 billion active daily users in the rest of the world). By failing to focus on the situation for those users who live outside the United States and E.U., the risk is that the literature is settling on what statisticians would describe as an underspecified model—one that fails to factor in important variables that are present in the reality it seeks to reflect. *See* W. Paul Vogt, Dictionary of Statistics & Methodology 332 (2011) (defining underspecified model as "[a] causal model in which important variables have been omitted."). With thanks to Susan Franck for suggesting the applicability of this terminology.

[24] *See, e.g.*, Sabelo Mhlambi et al., Think South: Reimagining the Internet (2019), https://archive.org/details/think_south_2019/ [https://perma.cc/4EUL-3EYH]; Nick Couldry & Ulises A.

doing important work to diversify the scholarly accounts of online governance in the U.S. context.[25] Legal scholars from the Global South have integrated some of these observations into their work on domestic regulation,[26] and the CyberBRICS project, hosted out of the Fundação Getulio Vargas Law School in Rio de Janeiro, has recently begun to work through issues of significance for major Global South powers of Brazil, Russia, India, China, and South Africa.[27] By bringing a wider range of case studies into conversation with the triadic literature for the first time, this Article builds on their work to develop a more complete and broadly applicable account of how online content is regulated.

Social media platforms varying in size and scope exist in many parts of the world,[28] but the major social media platforms with global reach—Facebook, Twitter, YouTube—are owned by U.S. companies and headquartered in Silicon Valley.[29] Their founders were American males in their twen-

Mejias, Costs of Connection: How Data Is Colonizing Human Life and Appropriating It for Capitalism (2019); Nanjala Nyabola, Digital Democracy Analogue Politics 31–48 (2018); Margaret E. Roberts, Censored: Distraction and Diversion Inside China's Great Firewall (2018); Zeynep Tufekci, Twitter and Tear Gas: The Power and Fragility of Networked Protest (2017); Rebecca MacKinnon, Consent of the Networked: The Worldwide Struggle for Internet Freedom (2012). For essential resources in this field, *see generally* Global Voices, https://globalvoices.org [https://perma.cc/Y5EQ-2U7Z] (last visited Oct. 23, 2020); Data & Soc'y, https://datasociety.net [https://perma.cc/M3HG-QL5N] (last visited Oct. 23, 2020).

[25] *See, e.g.*, LaTanya Sweeney, *Discrimination in Online Ad Delivery*, 56 Comms. ACM 44 (2013); Ruha Benjamin, Race After Technology (2019).

[26] *See, e.g.*, Chinmayi Arun et al., Hate Speech Laws in India (2018), https://drive.google.com/file/d/1pDoIwlusnM3ys-1GAYbnTPmepU22b2Zr/view [https://perma.cc/Q46C-PG95] (comprehensive overview of the way in which hate speech laws in India intersect with regulation by online platforms); Daniel Sive & Alistair Price, *Regulating Expression on Social Media,* 136 S. Afr. L.J. 51 (2019) (describing both direct and indirect strategies the South African government uses to control social media); Rasul Oriyomi Olukolu, Abigail Odozi Ogwezzy-Ndisika, Babatunde Adeshina Faustino & Faith Aanu Oloruntoba, *Social Media, Regulation and Challenges of Communication in an Evolving Nigerian Society*, 7 U. Balt. J. Media L. & Ethics 68 (2019) (theorizing the Nigerian government's relationship to social media).

[27] *See, e.g.*, Anja Kovacs, *Data Sovereignty, of Whom? Limits and Suitability of Sovereignty Frameworks for Data in India*, CyberBrics (Sept. 21, 2020), https://cyberbrics.info/data-sovereignty-of-whom-limits-and-suitability-of-sovereignty-frameworks-for-data-in-india [https://perma.cc/3G3Y-LBE5]; Torsha Sarkar, *Rethinking the Intermediary Liability Regime in India*, CyberBrics (Aug. 12, 2019), https://cyberbrics.info/rethinking-the-intermediary-liability-regime-in-india/ [https://perma.cc/5WQX-DJTB]; Sanilo Donesda & Diego Machado, *Smart Cities, Personal Data and Citizens' Rights in Brazil*, CyberBrics (May 22, 2019), https://cyberbrics.info/smart-cities-personal-data-and-citizens-rights-in-brazil/ [https://perma.cc/82YG-9K2A].

[28] Asha Barbaschow, *WeChat Sets the Record Straight for Its 690,000 Aussie Users: The Company Says WeChat is Not Governed by Chinese Law*, ZDNet (Oct. 1, 2020), https://www.zdnet.com/article/wechat-sets-the-record-straight-for-its-690000-aussie-users/ [https://perma.cc/9BYE-VT96] (WhatsApp has two billion active users, Douyin and Kuaishou each have 400 million active users in China, and Sina Weibo has 550 million active users in China); Simon Kemp, *More Than Half of the People on Earth Now Use Social Media*, Hootsuite (Aug. 10, 2020), https://blog.hootsuite.com/simon-kemp-social-media [https://perma.cc/3TUM-V9LQ]; *Lithuanian IT Entrepreneurs Find Success in Africa*, Lith. Trib. (May 26, 2016), https://lithuaniatribune.com/lithuanian-it-entrepreneurs-find-success-in-africa/ [https://perma.cc/6AH5-PR72] (Eskimi has 21 million users, mostly in Africa).

[29] Note that I am not assuming these particular companies, nor U.S. companies in general, will dominate this space forever. MySpace provides a useful reminder of the temporal contingency of the online market. I am also aware that one could argue at the margins over which companies to focus on in

ties, predominantly white and college educated.[30] The cultural and social footprint of the founders was built into the design of the platforms. "[T]he platforms we use were originally designed for small, specific, and incredibly mono-cultural communities . . . It was a predominately white and male space . . . It wasn't designed for everyone."[31]

As the platforms grew beyond the fairly homogenous initial set of American users, they came under increasing pressure to develop their content moderation model.[32] They strengthened systems that would allow users to identify material they found offensive, and built content moderation teams to develop policies that would both respond to particular cases and proactively address systemic clashes over cultural and social norms. With the development of geo-blocking tools, they also gained the ability to follow local law within territorially defined boundaries.[33]

Still, the leadership and senior staff of the platforms have, for the most part, remained socially and culturally homogenous. It is not simply that they are American, but that to this day, they represent a very narrow slice of the U.S. demographic. Facebook's latest diversity report, for example, detailed that women make up less than one-third of their senior leadership, and that African-Americans constitute a breathtakingly low 3.4% of the senior leadership.[34]

The impact of this homogeneity is reflected at every level of the platforms' operations. To take a concrete example, it is standard practice for the

this given moment. The three I highlight have been the focus of recent scholarship and have the benefit of a long (in internet years) lifespan of operations to draw from. But one could well have added, say, Instagram—now owned by Facebook—to the list. Lastly, this Article does not take on what are increasingly pressing questions about the governance of social media sites on the Dark Web. For an interesting ethnography on this topic, *see* Robert W. Gehl, *Power/Freedom on the Dark Web: A Digital Ethnography of the Dark Web Social Network*, 18 New Media & Soc'y 1219 (2016).

[30] Nicholas Carlson, *The Real History of Twitter,* Bus. Insider (Apr. 13, 2011, 1:30 PM), https://www.businessinsider.com/how-twitter-was-founded-2011-4?op=1 [https://perma.cc/4FBX-QASG]; Nicholas Carlson, *At Last: The Full Story of How Facebook Was Founded,* Bus. Insider (Mar. 4, 2010, 4:10 AM), https://www.businessinsider.com/how-facebook-was-founded-2010-3 [https://perma.cc/CRE2-PZCV]; Megan Rose Dickey, *The 22 Key Turning Points in the History of YouTube,* Bus. Insider (Feb. 15, 2013, 9:01 AM), https://www.businessinsider.com/key-turning-points-history-of-youtube-2013-2 [https://perma.cc/DML6-KV72].

[31] Caroline Sinders, *Designing a Better Internet,* British Council, https://www.britishcouncil.org/anyone-anywhere/explore/digital-identities/better-internet [https://perma.cc/DEQ7-62FS] (last visited Oct. 7, 2020).

[32] Klonick, *supra* note 16, at 1633; Julian Dibbel, *A Rape in Cyberspace*, Village Voice (Oct. 18, 2005), https://www.villagevoice.com/2005/10/18/a-rape-in-cyberspace [https://perma.cc/N7PU-A4JM] (spurring a wealth of discussion about the role of norms and rules in the regulation of online environments). *See Developments in the Law: The Law of Cyberspace*, 112 Harv. L. Rev. 1574, 1592–93 (1999).

[33] *Seeinfra* Part II.

[34] *See Diversity*, Facebook, https://diversity.fb.com/read-report/ [https://perma.cc/EGC9-G2HH] (last visited Oct. 7, 2020). The 2019 data also notes that Hispanic people make up only 3.5% of the senior leadership positions. *See also* Klonick, *supra* note 16, at 1621 ("A common theme exists in all three of these platforms' histories: American lawyers trained and acculturated in American free speech norms and First Amendment law oversaw the development of company content-moderation policy."); Tarleton Gillespie, *Governance of and by Platforms*, *in* SAGE Handbook of Social Media (Jean Burgess, Alice Marwick & Thomas Poell eds., 2017).

engineering teams at these platforms to work up new product designs by generating user personas. Akin to a baseball trading card, these user personas list key data points about hypothetical future users.[35] But as Caroline Sinders—a product designer who has consulted for the major social media platforms—explains, these user personas are typically constructed from a "benign Western policy perspective."[36] Thus there may be a user persona for Steve, a retiree in Florida who uses Facebook to stay connected with his grandchildren in Arizona, but there is unlikely to be a persona to forecast how Khalid, a state-affiliated actor who creates fake content attacking critics of Saudi Prince Mohammad bin Salman, would use the product.[37] Nor is there a persona to foretell how Bayarmaa, a female journalist in Mongolia, would be affected by the product design.[38]

It is not just the cultural orientation of the Silicon Valley-based platforms that draws them to default toward certain locations: the financial incentives also skew attention toward the West. For example, Facebook's earnings statement for the second quarter of 2020 indicates that average advertising revenue per user based in the United States and Canada was US$36.49. The same figure for users in the Asia Pacific was $2.99 and for users in the "rest of the world" category (those in the Middle East and Africa), the number was just $1.78.[39] In other words, a user in the United States is, on average, more than eleven times more valuable to Facebook's bottom line than a user in Myanmar, and more than sixteen times more valuable than a user in South Sudan.

As the case studies in this Article show, the content moderation systems these platforms designed have often failed users in the Global South as well as millions of users from more diverse backgrounds across the Global North.[40] These case studies elucidate the limited nature of the assumptions that underlie the platforms' content moderation systems by focusing on places that U.S. legal scholars have largely ignored, drawing attention to the

[35] *See, e.g.*, Tony Ho Tran, *Five Essentials for Your User Persona Template (with Examples),* InVision: Inside Design (Mar. 22, 2019), https://www.invisionapp.com/inside-design/user-persona-template/#What%20is%20a%20user%20persona [https://perma.cc/KR5A-X3SJ].

[36] Interview with Caroline Sinders in Tunis, Tunisia (June 14, 2019).

[37] *See* Manal Al-Sharif, *The Dangers of Digital Activism,* N.Y. Times (Sept. 16, 2018), https://www.nytimes.com/2018/09/16/opinion/politics/the-dangers-of-digital-activism.html [https://perma.cc/RG26-85BG].

[38] *See* Globe International Center, *Appeal for Justice in the Killing of Journalist in Mongolia*, ifex (Mar. 9, 2016), https://ifex.org/appeal-for-justice-in-the-killing-of-journalist-in-mongolia [https://perma.cc/G7PK-7SVZ].

[39] Presentation, Facebook, Facebook Q2 2020 Results (July 30, 2020) (available at https://s21.q4cdn.com/399680738/files/doc_financials/2020/q2/Q2-2020-FB-Earnings-Presentation.pdf [https://perma.cc/9X95-LUNE]).

[40] Jessica Guynn, *Facebook While Black: Users Call it Getting 'Zucked,' Say Talking About Racism is Censored as Hate Speech,* USA Today (Apr. 24, 2019, 7:26 AM), https://www.usatoday.com/story/news/2019/04/24/facebook-while-black-zucked-users-say-they-get-blocked-racism-discussion/2859593002 [https://perma.cc/EHV7-AE55].

life-threatening impact that the failure to see the contingency of these assumptions has had and is having on millions of users.

Take, for example, Myanmar, where members of the Rohingya Muslim minority have recently faced genocide. According to a U.N. factfinding mission, beginning in 2012, Myanmar's military personnel developed a sophisticated disinformation campaign on social media, using fake accounts attracting millions of followers, to systematically spread hate and incite violence against the Rohingya people.[41] Facebook failed to understand how its platform was being weaponized in Myanmar before thousands were killed and millions more displaced.[42] This is, in part, because when Facebook launched in Myanmar, the company's content moderation policies focused on the United States and the E.U. where, at that time, private actors were the main source of troubling online activity.[43] The tragedy in Myanmar can also be attributed to Facebook's total failure to do basic due diligence or develop the cultural competence needed to responsibly enter the Myanmar market in advance of launching its platform there.

There are no Facebook content moderators in Myanmar and, until recently, none outside of Myanmar who spoke Burmese.[44] Facebook relied on ordinary users in Myanmar to flag offensive or hateful material. Such a flagging system operates on the assumption that offensive content is created and tolerated by only a minority of an online community. This works for the kind of contexts that the platform founders had in mind. For example, when child abuse content appears online there is a good chance people will report it because most people do not endorse child abuse. Unfortunately, persecution of the Rohingya is tolerated across all parts of Burmese society.[45] In other words, expecting Burmese users to alert Facebook to incitement against the Rohingya in Myanmar would be like relying on the majority Hutu population to call out incitement against the Tutsi in 1994 Rwanda.[46]

This Article's first contribution is to introduce a new set of case studies into the triadic literature. Consistent with leading accounts in the scholar-

[41] Human Rights Council, Rep. of the Detailed Findings of the Indep. Int'l Fact-Finding Mission on Myan., U.N. Doc. A/HRC/39/CRP.2 (2018) [hereinafter Myan. FFM Rep.].

[42] Steve Stecklow, *Inside Facebook's Myanmar Operation Hatebook: A Reuter's Special Report,* Reuters (Aug. 15, 2018, 3:00 PM), https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/ [https://perma.cc/HE6J-AWD8].

[43] *See* Tim Wu, *Is the First Amendment Obsolete?,* 117 Mich. L. Rev. 547, 557 (2018) (explaining that in the existing literature, examples of inciting speech "were always presumed to be malicious private actors: the missing piece was the threat posed by state-organized trolling and systematic use of hate speech").

[44] *See infra* Part II.A.2.

[45] *See* Myan. FFM Rep., *supra* note 41, ¶¶ 20–23.

[46] Or, for that matter, like relying on the 8chan community to call out incitement against the Hispanic community in 2019 America. *See* Gianluca Mezzofiore & Donie O'Sullivan, *El Paso Mass Shooting Is at Least the Third Atrocity Linked to 8chan This Year*, CNN (Aug. 5, 2019, 7:43 AM), https://www.cnn.com/2019/08/04/business/el-paso-shooting-8chan-biz/index.html [https://perma.cc/3T4A-SBRZ].

ship, this intervention is framed in the context of a pluralist system of online speech governance.[47] Taken seriously, the methodological orientation of pluralism, with its focus on empirical observation, opens up a new set of questions that can emancipate the triadic literature from the limiting assumptions that have started to become embedded within it. This provides the backdrop for the second contribution of this Article, which is to strengthen the free speech triangle's analytic power by focusing on what the actors in the triangle actually do in order for content to reach the global public square. The result is what one might term a function-based approach, organized around content production, content amplification, rule creation, and enforcement.

In making this turn toward function, this Article does not suggest that formal status—for example, whether one is a state, individual, or corporation—is unimportant. The fact that a state is an actor with public law responsibilities matters regardless of what activities that state undertakes. Likewise, the status of technology companies as private actors is important to keep in mind when considering their obligations. However, as the triadic literature shows, the exclusive use of a status-based approach invariably leads us to miss the diversity of roles that various actors take on. More significantly, this approach risks embedding status-based assumptions that deny the lived experience of millions of social media users worldwide.

Before turning to the structure of the Article, a number of clarifications are in order. First, while the term "global public square" has rhetorical value, a more accurate title would be *Governing the Global Public Square(s)* as the internet simultaneously supports a global public square (singular) and a multiplicity of global public squares (plural).[48]

The longstanding commitment of internet technologists to perfect interoperability means that, technically speaking, there is a single global public square.[49] Once people began communicating over the internet at scale, however, we rapidly created a multiplicity of squares. Some states created de jure boundaries of varying degrees—one can think of North Korea as an extreme example. For the most part though, linguistic, cultural, and social differences established de facto boundaries. The resulting public squares are

[47] "[T]he digital age features a pluralist model of speech control" wrote Jack Balkin in 2018; *see* Jack M. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation*, 51 U.C. Davis L. Rev. 1149, 1153 (2018). "As a first approximation, we might call this system triadic. . .[but]. . . the new system of speech regulation is actually far more complicated, because there are more than three sets of players." *Id.* at 1189.

[48] Or, as one of Twitter's policy leads put it in a quote that captures the simultaneously global and local dimensions of their activity: "We get one shot to write a policy that has to work for 350 million people who speak 43-plus languages while respecting cultural norms and local laws." Jerrel Peterson, *quoted in*, Kate Conger, *Twitter Backs Off Broad Limits on 'Dehumanizing' Speech*, N.Y. Times (Jul. 9, 2009), https://www.nytimes.com/2019/07/09/technology/twitter-ban-speech-dehumanizing.html [https://perma.cc/4K69-PCTF]. I am grateful to David Pozen for helping me clarify my thinking on this point.

[49] *See generally* Milton L. Mueller, Networks and States: The Global Politics of Internet Governance (2010).

still largely global—access to them is not territorially bounded nor defined by nationality—but there is nothing inherently universal about them. While theoretically open to much of the global public, the regulatory dynamics of the online space for the local and diaspora populations of Abuja, Nigeria, are unlikely to be the same as those in the space focused on Dayton, Ohio. Thus, while this article uses the term global public square in the singular to avoid the awkwardness of a pluralizing parenthetical, readers should remain cognizant of the tension that policymakers in both governments and platforms face as they navigate the regulation of space that is at once localized and globalized. In other words, those making determinations about what speech should stay online and what should be removed need to be aware of the universal dimension of their actions, even as they react to hyper-local developments.

Next, there are two important points in relation to the case studies presented here. First, while providing a significant expansion of the material currently found in the triadic literature, these case studies still provide only a partial view of activities happening across the globe. Privacy constraints and the newness of transparency reporting by platforms means that reports on content moderation decisions lack the granularity needed to build out a concrete example of governance activity around a given type of content.[50] Consequently, the countries highlighted in this article are those where the media and/or civil society has also been tracking content moderation decisions. Nonetheless, readers should remember that decisions (and non-decisions) about what content stays online are happening everywhere—including in places where the space for media investigation and civil society work is constrained or non-existent.[51] Second, this Article intentionally features case studies from the Global South as a counter to the prevalence of examples from the United States and the E.U. in the current literature.[52] However, this decision does *not* imply that the dynamics uncovered in these new case studies are only applicable to the Global South. Many of the insights that these case studies reveal are applicable in understudied examples concerning marginalized populations throughout the Global North.

Part I begins by briefly describing the mechanics of content moderation; then it proceeds to sketch out recent developments that have captured the

[50] *See, e.g.*, *Facebook Transparency Report*, FACEBOOK, https://transparency.facebook.com/ [https://perma.cc/39TD-MQTF] (last visited Oct. 7, 2020).

[51] For a useful overview of the global narrowing civil society space, see BERNSTEIN INSTITUTE FOR HUMAN RIGHTS, DEFENDING DISSENT: CIVIL SOCIETY AND HUMAN RIGHTS IN THE GLOBAL CRACKDOWN (2017), http://www.law.nyu.edu/sites/default/files/Bernstein%20Institute%20Conference%20Digest.pdf [https://perma.cc/DP6M-VJ42].

[52] In introducing multiple case studies from the Global South, I am grateful to Chinmaya Arun for her insight that such an approach can help "to show that there are many ways in which Southern populations are affected by technology." Chinmaya Arun, *AI and the Global South: Designing for Other Worlds in* OXFORD HANDBOOK OF ETHICS OF AI, 590, 591 (Dubber et al., eds. 2020). Just as the "global public square" is, in fact, a multiplicity of squares, the "Global South" is likewise a term that captures a plural reality.

attention of prominent legal scholars and provided the empirical backdrop for their work. These developments arise primarily from the United States and the E.U., although several scholars also reference decisions that different technology companies have made regarding the Chinese government's demand that they censor material as a condition of entry into the Chinese market.[53] Finally, Part I describes key insights that have been derived from this work and offer an initial articulation of some of the assumptions embedded within it.

Part II draws on public sources, background interviews, and technology company transparency data to introduce a new set of case studies and provide context for illustrative debates over the regulation of online speech in Myanmar, India, Sri Lanka, South Sudan, and Turkey. These case studies reveal the non-universality of key assumptions that underlie the platforms' content moderation systems. Approaching the global public square as a site of pluralism, the case studies show that online speech regulation is far more complex than the triadic literature has thus far allowed. Case studies of the Global South highlight diverse functions of, and relationships between, users, civil society groups, traditional media, international organizations, governments, courts, and platforms that are minimized or absent in existing work. Through these case studies we see that these actors can each take on a range of roles: States and technology companies are not just regulators, they also create and amplify content; users do not just post material, they also play a regulatory role. The case studies also emphasize the vital role that linguistic and cultural context plays in how the global public square functions.

Part III proposes supplementing the free speech triangle with a function-based approach to enhance our understanding of the process that controls what appears or does not appear on our screens. The four points on the function-based model—content production, content amplification, rule creation, and enforcement—provide a useful framework for structuring the way we think about the regulation of online space without locking us into any parochial assumptions about who might take on which function(s). It embraces, rather than sidesteps, the pluralism inherent in conceptions of the in/appropriateness of different types of speech across and within diverse communities around the world, and in so doing it provides a roadmap for policymakers interested in identifying parts of the system that are ripe for regulatory action.

[53] *See infra* Part I.B.

I. The triadic literature

The first major pieces of legal scholarship on the regulation of cyberspace appeared in the late 1990s.[54] The academics in conversation at that time could be roughly divided into those who believed that the internet would usurp the regulatory role of states, and those who did not.[55] The state-user dyad was central to this early work, but then along came social media.[56]

A 2006 article by Seth Kreimer observed that "in contrast to the usual free expression scenario, the internet is not dyadic."[57] Articles by Rebecca Tushnet and Jack Balkin in the late 2000s contained robust discussions of the role of technology companies in relation to free expression.[58] Julie Cohen's work on informational capitalism brought a political economy lens to the analysis of technology companies.[59] A 2011 article by Danielle Keats Citron and Helen Norton was the first to foreground the role of technology companies in the regulation of online hate speech,[60] and in 2014, Marvin Ammori highlighted the significance of technology companies in free speech lawyering more generally.[61]

The conceptual shift to a triadic model of speech regulation came with Balkin's 2018 essay, "Free Speech is a Triangle,"[62] and Kate Klonick's 2018 article, "The New Governors."[63] Although Balkin highlighted the

[54] *See* Lawrence Lessig, Code and Other Laws of Cyberspace (1999). This was the leading work of this era.

[55] John Perry Barlow, *supra* note 12 (proclaiming the freedom of the internet from state control); Mike Godwin, Cyber Rights: Defending Free Speech in the Digital Age 44, 77–78 (2003); *cf.* Goldsmith & Wu, *supra* note 13, at 180–84 (2006).

[56] Facebook launched in 2004, YouTube in 2005, and Twitter in 2006. Here Comes Everybody: The Power of Organizing Without Organizations by NYU Professor Clay Shirky was one of the first major books to foreground the role of social media platforms in the online space. *See* Clay Shirky, Here Comes Everybody: The Power of Organizing Without Organizations (2008).

[57] Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. Pa. L. Rev. 11, 14 (2006). Interestingly though, Lessig's 1999 work foreshadowed exactly this insight with his insight that "code is law." However, scholars did not focus on the implications of this insight until much later.

[58] Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 Geo. Wash. L. Rev. 986 (2008); Jack M. Balkin, *The Future of Free Expression in a Digital Age*, 36 Pepp. L. Rev. 427, 432–35 (2009).

[59] *See generally* Julie E. Cohen, Between Truth and Power: The Legal Constructions of Informational Capitalism (2019). It is worth noting that Cohen's work in this space extends back much earlier.

[60] Danielle K. Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age*, 91 B.U. L. Rev. 1435, 1436 (2011). *See also* Danielle K. Citron & Mary A. Franks, *Criminalizing Revenge Porn*, 49 Wake Forest L. Rev. 345 (2014).

[61] Marvin Ammori, *The "New" New York Times: Free Speech Lawyering in the Age of Google and Twitter*, 127 Harv. L. Rev. 2259 (2014). *See also* Kyle Langvardt, *Regulating Online Content Moderation*, 106 Geo. L.J. 1353, 1358 (2018) ("the twentieth century's law of free expression established that only the final censor—at that time, the state—was subject to law. Today, a small number of politically-unaccountable technology oligarchs exercise state-like censorship powers without any similar limitation.").

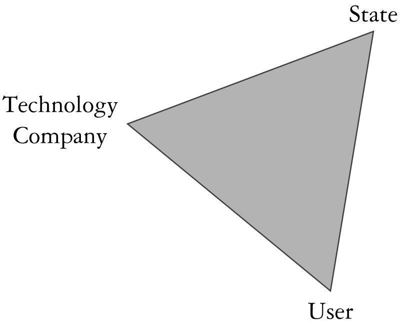
[62] Balkin, *supra* note 16. The essay built on themes developed in his 2014 Harvard Law Review article. *See* Balkin, *supra* note 19.

[63] Klonick, *supra* note 16, at 1664 (describing "a new model of free expression: a triadic model. In this new model, online speech platforms sit between the state and speakers and publishers.").

inherent *plurality* of the system,[64] this observation has thus far been minimized as scholars have begun to extend the triadic approach to issues of censorship, surveillance, and data control.[65]

In a 2019 article, Kristen Eichensehr built out the triadic model, presenting triangular schematics to forecast the position of technology companies relative to governments and users in a range of scenarios.[66]

FIGURE 1: STATUS-BASED TRIAD



The figure above, adapted from Eichensehr's 2019 article, is a simplified schematic of the free speech triangle introduced in Balkin's 2018 essay. The innovation of the free speech triangle is to add technology companies into the set of relevant actors involved in determining what content gets to appear and stay online.

The types of actors identified in the schematic are crucial in understanding how the global public square operates. However, because of the limited set of examples the model was developed from, and has thus far been applied to, the triadic literature has started to lock in a set of non-generalizable assumptions about the roles that each of these actors play.

The following section sets the stage by briefly explaining the basic mechanics of content moderation. It then looks at the real-world developments that the triadic literature has drawn from, observing a common set of case studies arising from the United States and the E.U. It highlights the key insights yielded from this scholarship, and concludes by foreshadowing

[64] Balkin, *supra* note 47, at 1186 (includes a subsection entitled "From the Dyadic to the Pluralist Model of Speech Governance").

[65] *See* Danielle K. Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035 (2018) (role of technology companies in content moderation); Alan Z. Rozenshtein, *Surveillance Intermediaries*, 70 STAN. L. REV. 99 (2018) (role of technology companies in surveillance); Daskal, *supra* note 14 (role of technology companies in data collection); Bloch-Wehba, *supra* note 17 (role of technology companies in content moderation).

[66] Eichensehr, *supra* note 16, at 704 (predicting that "companies will fight against or resist governments when the companies perceive themselves to be and can credibly argue that they are protecting the interests of users against governments").

assumptions embedded within it that will become clearly apparent through the case studies in Part II.

A. *The Mechanics of Content Moderation*

Major social media platforms use a mix of human beings (employees, contractors, and unpaid users) and artificial intelligence ("AI"), to bar content that violates their internal standards. The platforms screen out some content before it ever appears online and remove other content on a timeline ranging from seconds to years after it appears.[67]

Based in the United States, the major social media platforms operate within a highly permissive legal environment regarding content moderation.[68] The environment is permissive because, in passing Section 230 of the 1996 Communications Decency Act ("CDA"), Congress sought to ensure that technology companies would not be deterred from moderating content that appeared on their sites.[69] Section 230 stops the platforms from being treated as either publishers or speakers of user content, meaning they are immunized from liability both when they fail to remove unlawful content and when they do remove lawful content.[70] Recent content moderation decisions by many technology companies have, however, spurred both President Trump and President-elect Biden to revive a push to limit liability protection under Section 230.[71]

As a strictly legal matter, there is no reason for the platforms to have developed the elaborate content moderations systems they currently run. However, just like ungoverned spaces in the offline world, platforms quickly become unsafe and ultimately "uninhabitable" without moderation.[72] For their business model to function, platforms need to keep users online so that they can be exposed to the advertising that generates the technology compa-

[67] *See generally* TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET (2018).

[68] Anupam Chander, *Internet Intermediaries as Platforms for Expression and Innovation* 1, 6 (Glob. Comm'n on Internet Governance, Paper No. 42, 2016) ("While many European and Asian nations leave intermediaries open to liability for the actions of their users in certain cases, the United States generally limits liability").

[69] For a fascinating legislative history of the CDA, *see* Robert Cannon, *The Legislative History of Senator Exon's Communications Decency Act: Regulating Barbarians on the Information Superhighway,* 49 FED. COMM. L.J. 52 (1996).

[70] 47 U.S.C § 230(c)(1)–(2) (2019).

[71] *See* Nellie Bowles, *The Complex Debate Over Silicon Valley's Embrace of Content Moderation*, N.Y. TIMES (June 5, 2020), https://www.nytimes.com/2020/06/05/technology/twitter-trump-facebook-moderation.html [https://perma.cc/H9D6-PTGX] (discussing the recent increased moderation, Trump's executive order, and Biden's comments on reform.); *see also* Davey Alba et al., *supra* note 4 (describing Twitter's inconsistent moderation of false or incendiary tweets from politicians including the first fact-check notices placed on the U.S. president's often misleading tweets and a subsequent warning—the first of its kind for any public figure—that a tweet glorified violence); Zakrzewski, *supra* note 10, (discussing the same and the selective removal of misleading tweets about COVID-19 by Brazil's President Jair Bolsonaro and Venezuela's President Nicolás Maduro).

[72] *See* James Grimmelmann, *The Virtues of Moderation*, 17 YALE J.L. & TECH. 42 (2015).

nies' revenue.[73] To keep users online—and thus advertising revenue flowing—platforms had to develop standards for what would and would not be acceptable. Without the strictures of formal law to bind them, their first efforts in this regard were driven by instinct. In the words of a former Facebook content moderator, "if it makes you feel bad in your gut, then go ahead and take it down."[74]

As Klonick documented in her study, Facebook, Twitter, and YouTube increasingly formalized their approach to content moderation over time.[75] Guided by the First Amendment principles that the founders of these platforms were familiar with, what began as a one-page list of standards eventually become an eighty-page rule book.[76] In addition to these internal rules, the platforms began sharing summaries of these rules with the public in the form of globally applicable "Community Standards."[77] Facebook's current set of Community Standards runs some twenty pages, with guidance such as "Do not post: Content that depicts real people and laughs at or makes fun of their implied or premature death, serious or fatal disease or disability, nonconsensual sexual touching, domestic violence, or serious physical injury."[78] Both the internal rules and the public-facing Community Standards at the different platforms are regularly updated to reflect how the platforms' concerns evolve over time.[79]

In terms of the enforcement of these standards, it is crucial to first understand the scale of content that the platforms are dealing with: 6,000 tweets per second; over 500 hours of YouTube video uploaded every minute; 350 million Facebook photos uploaded each day.[80] This scale has necessitated a growing industry of commercial content moderators. Some are Facebook employees, but most of those who do the trauma-inducing work of keeping the darkest parts of humanity off our news feeds are contract workers on a

[73] Facebook*, supra* note 23 (as of the second quarter of 2020, advertising income comprised over 98% of Facebook's revenue).

[74] Charlotte Willner, *quoted in* Klonick, *supra* note 16, at 1631.

[75] Klonick, *supra* note 16, at 1632.

[76] *Id.* at 1631–35.

[77] *Id.* at 1642. The platforms use different terms for these public-facing standards. "Community Standards" is the term used by Facebook. Twitter uses "Rules" and YouTube uses "Community Guidelines." For simplicity going forward, I use Community Standards when referencing this general category of public-facing standards.

[78] *Community Standards*, Facebook, https://www.facebook.com/communitystandards/cruel_insensi tive [https://perma.cc/EVC2-JQNA] (last visited Oct. 7, 2020).

[79] *See* Catherine Bruni & Somraya Chamali, *The Secret Rules of the Internet: The Murky History of Moderation, and How It's Shaping the Future of Free Speech*, Verge (Apr. 13, 2016), https://www.theverge.com/ 2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech [https://perma.cc/H9D6-PTGX] ("No booklet could ever be complete. No policy definitive.") (The fluidity of these rules is well captured through interviews with a former YouTube content moderator).

[80] Anmar Frangoul, *With Over One Billion Users, Here's How YouTube is Keeping Pace with Change*, CNBC (Mar. 14, 2018, 4:54 AM), https://www.cnbc.com/2018/03/14/with-over-1-billion-users-heres-how-youtube-is-keeping-pace-with-change.html [https://perma.cc/BX8U-5MQN]; *Twitter Usage Statistics*, Internet Live Stats, https://www.internetlivestats.com/twitter-statistics/ [https://perma.cc/ 7WZD-CCVW] (last visited Sept. 10, 2019); *Facebook by the Numbers*, Omnicore (Apr. 22, 2020), https:/ /www.omnicoreagency.com/facebook-statistics/ [https://perma.cc/UP5L-HVXW].

minimum wage.[81] Even then, these workers provide a fraction of the moderation the system demands.

One supplement to commercial content moderators is users themselves. Through a process known as "user flagging" platforms have set up mechanisms to allow ordinary users to report content they find offensive.[82] Another supplement comes in the form of AI; algorithms are trained to identify prohibited content.[83] However, despite the optimism accompanying discussions of future use of AI in content moderation generally, it is widely recognized that inherently contextual content, like hate speech, remains hard for any algorithm to identify with precision.[84] Finally, a technique known as geo-blocking enables the platform to prohibit specific content in certain geographic locations, even as that same content flows freely elsewhere.[85]

B. A Limited Dataset

Leading accounts of the regulation of online activity by American legal scholars are overwhelmingly centered on the United States and the E.U. Klonick's 2018 article provides the most diverse offering of examples through interviews with former lawyers at YouTube, Facebook, and Twitter, who recount difficult content moderation decisions regarding: a protest photo in Iran, a video of an Egyptian man being beaten, photos offensive to the Thai and Turkish governments, and the controversial video, *The Innocence of Muslims.*[86] Even so, Klonick focuses on high-stakes decision points in Silicon Valley, and largely abstracts these non-Western examples from their local context. Moreover, beyond the interview material in Klonick's article, there are only passing references to events outside the United States and the E.U. in the triadic literature.[87]

[81] For the most comprehensive study on this industry, *see* Sarah T. Roberts, BEHIND THE SCREEN: CONTENT MODERATION IN THE SHADOWS OF SOCIAL MEDIA (2019).

[82] *See* Gillespie, *supra* note 67, at 87–97.

[83] For a lucid explanation of this process, *see* Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1378 (2018).

[84] In U.S. congressional testimony, Facebook CEO Mark Zuckerberg acknowledged that in terms of areas ripe for AI solutions, "hate speech is one of the hardest, because determining if something is hate speech is very linguistically nuanced." Nonetheless, he went on to say, "Hate speech—I am optimistic that, over a 5 to 10-year period, we will have A.I. tools that can get into some of the nuances." Mark Zuckerberg, Remarks to the U.S. Senate Commerce and Judiciary Committees (Apr. 10, 2018), *in* WASH. POST, https://www.washingtonpost.com/news/the-switch/wp/2018/04/10/transcript-of-mark-zuckerbergs-senate-hearing/?utm_term=.69a8fb76815d [https://perma.cc/Z92T-DCGE].

[85] *See* Karl Schaffarczyk, *Explainer: What is Geoblocking?*, CONVERSATION (Apr. 17, 2013, 12:31 AM), https://theconversation.com/explainer-what-is-geoblocking-13057 [https://perma.cc/XGT6-WBNF].

[86] *See* Klonick, *supra* note 16, at 1619–25.

[87] *See, e.g.*, Balkin, *supra* note 19, at 2308 n. 37 (referencing the Egyptian government temporarily shutting down the internet during the 2011 Arab Spring protests); Jack Goldsmith, *The Failure of Internet Freedom,* KNIGHT FIRST AMEND. INST. (June 13, 2018), https://knightcolumbia.org/content/failure-internet-freedom [https://perma.cc/H9D6-PTGX] (A section entitled "Failure Abroad" focuses primarily on China and the EU, but includes one paragraph on authoritarian Arab states, noting that they have "grown adept at employing internet shutdowns and slowdowns, at blocking encrypted communication

With some frequency, scholars cite statistics that place the vast majority of users outside the United States, even as they focus their work on developments inside America.[88] Interactions between the U.S. Congress, the U.S. Supreme Court, and technology companies over how to restrict child pornography; events flowing from the 2010 Wikileaks disclosures of diplomatic cables; and decision-making over content removal in relation to the 2017 "Unite the Right" rally in Charlottesville, receive recurrent scrutiny.[89] Scrutiny will likely continue to focus on platform reactions to trends within the United States, including the current Black Lives Matter protests, U.S. presidential and congressional tweets, and the discourse around the country's 2020 election.[90]

Efforts to look abroad focus mainly on the E.U., with the regulation of hate speech and the enforcement of data privacy laws as the primary issues of concern.[91] In a notable deviation from the norm, Jennifer Daskal highlights and justifies her decision to focus on case studies exclusively from the

tools, and at cracking down on circumvention efforts."); Eichensehr, *supra* note 16, at 674 (includes a footnote describing Canadian company Blackberry's decision to withdraw from the Pakistani market).

[88] *See, e.g.*, Marvin Ammori, *The "New" New York Times: Free Speech Lawyering in the Age of Google and Twitter*, 127 HARV. L. REV. 2259, 2263 (2014); Eichensehr, *supra* note 15, at 668.

[89] *See, e.g.,* Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 71–77 (2006) (discussing restrictions on child pornography); Rebecca Tushnet, *Power Without Responsibility: Intermediaries and the First Amendment*, 76 GEO. WASH. L. REV. 986, 996–1002 (2008) (same); Bambauer, *supra* note 17, at 863, 891–94; Benkler, *supra* note 17 (discussing Wikileaks); Balkin, *supra* note 19, at 2327–29; Balkin, *supra* note 15, at 2013 (discussing Charlottesville); Kyle Langvardt, *Regulating Online Content Moderation*, 106 GEO. L.J. 1353, 1358 (2018). Although outside this Article's focus on speech regulation, the interaction between the FBI and Apple following the San Bernardino shooting is another development that receives repeated attention from those working within a triadic approach. *See, e.g.,* Eichensehr, *supra* note 16, at 678, 698.

[90] *See* Hadas Gold, *Facebook Will Start Labeling Pages and Posts from State-Controlled Media*, CNN BUS. (June 5, 2020, 5:54 AM), https://www.cnn.com/2020/06/04/media/facebook-state-media-label/index.html [https://perma.cc/9VPA-GTUQ] (detailing that media outlets are scrutinizing Facebook's timing of the rollout of labeling state-controlled media posts and official pages and geo-blocking state-run media advertisements within the United States); *see also* Cat Zakrzewski, *The Technology 202: President Trump's Conspiracy Theories Are Putting Pressure on Twitter to Change Its Policies,* WASH. POST (June 10, 2020, 9:40 AM), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/10/the-technology-202-president-trump-s-conspiracy-theories-are-putting-pressure-on-twitter-to-change-its-policies/5edfe562602ff12947e888d1 [https://perma.cc/M26C-5TCW] (discussing Trump's sharing of conspiracy theories); *see also* Kate Conger, *Twitter Places Warning on Florida Congressman's Tweet*, N.Y. TIMES (June 1, 2020), https://www.nytimes.com/2020/06/01/technology/twitter-matt-gaetz-warning.html [https://perma.cc/44GS-A5BU] (describing the labeling of a congressman's tweet that prevented it from being "retweeted or liked, to prevent it from being amplified."). *But see* Henry Olsen, Opinion, *Twitter's Double Standard Has Become Painfully Clear*, WASH. POST (May 29, 2020, 4:06 PM), https://www.washingtonpost.com/opinions/2020/05/29/twitters-double-standard-has-become-painfully-clear/ [https://perma.cc/N63W-VA6L] (decrying that American conservative politicians have faced twitter censorship but Iran's Ayatollah Khamenei's tweets about Israel remain without any label or censorship).

[91] *See, e.g.,* Citron, *supra* note 65 (E.U. hate speech regulation); Daskal, *supra* note 14, at 185 (E.U. data privacy and the right to be forgotten); Balkin, *supra* note 47, at 1206–07. For the latest developments in the right to be forgotten litigation, *see* Court of Justice of the European Union Press Release No. 112/19, Judgment in Case C-507/17, The Operator of a Search Engine is not Required to Carry out a De-referencing on All Versions of its Search Engine (Sept. 24, 2019).

United States and the E.U. in her 2018 article on data protection.[92] More commonly, this selection bias remains unacknowledged.

The only place that routinely receives attention outside the United States and the E.U. is China. However, with most social media platforms refusing to operate under the censorship conditions that the Chinese government imposes, the account is thin. Typically, scholars describe the Chinese government as a leading censor of content and recount various decisions by Google on whether to stay in the Chinese market.[93]

There are exceptions. In his 2011 *California Law Review* article "Googling Freedom," Anupam Chander takes a much deeper dive into Google's relationship to the Chinese government and Chinese users in order to consider the obligations of new media companies to those in unfree societies.[94] Outside of law journals, journalists and technology fellows have published a number of rich case studies.[95] Yet American legal scholars have largely been theorizing developments that affect the global populace by looking no further abroad than Europe.

Notwithstanding the limited set of examples from which the triadic literature was drawn, it has produced several important insights about the regulation of online speech. By studying the way technology companies have changed their Community Guidelines to prohibit hate speech, scholars have seen these companies shift away from the normative underpinnings of the First Amendment.[96] By examining these companies' responses to the threat of fines for failure to comply with E.U. regulations, scholars strengthened the concept of "collateral censorship," where the state uses a private company to censor speech.[97] By recounting U.S. government efforts to limit the spread of diplomatic cables posted by Wikileaks, scholars have assessed the ways in which government can use soft power to secure its censorship (or, more neutrally, regulatory) goals.[98] Whether through hard or soft power, the triadic literature has highlighted the ways in which a collateral censorship

[92] *See* Daskal, *supra* note 14, at 185 ("[A]lthough I focus primarily on how these issues are playing out in the United States and European Union, I do not intend to suggest that these are the only important players . . . But the United States and European Union provide an important, interesting, and instructive place to start. They are large and powerful actors. And while they sometimes take divergent approaches, they share enough common values and normative assumptions that there is both the possibility and reality of increased harmonization in key areas. They thus make informative case studies.").

[93] *See, e.g.*, Kreimer, *supra* note 57, at 18; Ammori, *supra* note 61, at 2282–83; Eichensehr, *supra* note 16, at 674–78.

[94] Anupam Chander, *Googling Freedom*, 99 Cal. L. Rev. 1 (2011).

[95] *See generally supra* note 24.

[96] *See, e.g.,* Citron, *supra* note 65, at 1070 (recounting the "long history" of Silicon Valley embracing First Amendment norms, and the role of European regulation in pulling technology companies away from this baseline).

[97] Michael I. Meyerson, *Authors, Editors, and Uncommon Carriers: Identifying the "Speaker" Within the New Media*, 71 Notre Dame L. Rev. 79, 116–24 (1995).

[98] Summarizing the impact of soft power collateral censorship by the U.S. government in relation to Wikileaks, Derek Bambauer observes that "informal government pressures on key intermediaries accomplished what formal legal action likely could not." Bambauer, *supra* note 17, at 891–94; Balkin, *supra* note 19, at 2327–29.

relationship between government and technology companies can fuel due process violations against users.[99] Finally, by looking at the expansion of the E.U. Right to be Forgotten and data privacy rulings, scholars have observed how technology companies can spread local law extraterritorially.[100]

C. *That Which Remains Unseen*

The selection bias that has made primarily the United States and the E.U. most visible in the triadic literature has served to embed certain assumptions about the online regulation of speech globally. The non-universality of these assumptions is easy enough to miss until our gaze is expanded to a broader set of examples.

The triadic literature contains a default assumption that the state can play a regulatory role, because that is what the state does in the examples that the literature has focused on. The possibility of the state as an actor that cannot or will not regulate speech has been left unexplored. Likewise, the concerns about either collateral censorship or the global spread of local law arising from U.S. and European government efforts, assume a state that is able to command the attention of the relevant technology company[101]—another assumption that falls away with the case studies in Part II. Missing in all these accounts is the possibility of the state's primary activity being the use, rather than the regulation of social media.[102]

Less visible still is the assumption that any regulatory efforts made in the online space will play out against the backdrop of an offline environment with a functioning rule of law. Relatedly, the user-flagging model developed by the platforms assumes—and scholars have thus far not questioned—a world in which a minority of a given user population will engage in "dangerous speech,"[103] and that there will be users who are willing and able to "flag" such content for removal.

[99] This strand of the literature was catalyzed by Danielle Keats Citron's 2008 article, *Technological Due Process*, which critiqued the dearth of due process within technology systems used by U.S. government agencies. Danielle K. Citron, *Technological Due Process*, 85 WASH. U. L. REV. 1249 (2008). But the concern about a lack of due process has since been extended to content moderation decisions by technology companies. *See, e.g.,* Grimmelmann, *supra* note 72, at 66–67 (2015); Klonick, *supra* note 16, at 1668–69; Balkin, *supra* note 19, at 2031.

[100] *See, e.g.,* Daskal, *supra* note 14, at 185; Balkin, *supra* note 47, at 1206–07.

[101] In *Litigating Data Sovereignty,* Andrew K. Woods acknowledges that this may not always be the case. "Not all states can force a company to comply with their wishes, of course; just those with a big enough market to warrant the firm's attention." Andrew Woods, *Litigating Data Sovereignty,* 128 YALE L.J. 328, 405 (2018). He does not develop the point further, but nonetheless this is the only example of any such acknowledgement that I have found within the triadic literature.

[102] This may help explain the apparent absence of foresight – and delayed reactivity – on the part of both platforms and policymakers alike in terms of Russian interference in the 2016 U.S. presidential election. *See* MUELLER, *supra* note 22.

[103] For an explanation of this concept, *see* Susan Benesch, *What is Dangerous Speech?* DANGEROUS SPEECH PROJECT, https://dangerousspeech.org/about-dangerous-speech/ [https://perma.cc/7Q2Z-NUEC] (last visited Oct. 4, 2019) ("Dangerous Speech is any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group.")

The non-universality of these and other assumptions that the triadic literature has increasingly relied on will become apparent through the following case studies.

II. Expanding Our View

In the following Part, this Article introduces a new set of case studies into the literature about the regulation of online content. The case studies focus on controversial content moderation decisions (and non-decisions) arising from Myanmar, India, Sri Lanka, South Sudan, and Turkey. I conclude this Part by synthesizing key observations from the case studies that strengthen our understanding of how the global public square actually operates.

A. Case Studies from the Global South

The case studies that follow draw on the work of journalists and civil society groups, along with transparency data released by social media companies themselves. As a result of using these sources, the case studies focus on incidents that are viewed as problematic by at least one of the actors involved; neither journalists nor civil society groups spend their resources investigating non-issues. These case studies do not closely examine the thousands of content moderation decisions that are undertaken seamlessly each day, including within the locations examined in the case studies. However, the fact that online content regulation works much of the time, in many locations, does not detract from the urgency of attending to those instances where it does not work, especially when such instances can result in significant harm, or even death.

1. *Myanmar*

Recent atrocities incited by the Myanmar military against the minority Muslim population, the Rohingya, constitute the world's first genocide fueled by social media. This case study focuses exclusively on Facebook since it is the platform that Myanmar state officials used to incite atrocities. Moreover, virtually all internet users in Myanmar are on Facebook, while less than one percent of the population use Twitter or YouTube.[104] The following explains the context in which Facebook was introduced to Myanmar, highlights the failure of Facebook to set up a warning system that would work in the Myanmar context, and details some of the efforts made by civil society groups to fill this gap. It also shows how the efforts of civil society failed to garner Facebook's attention. Only a belated focus on the situation by Western media and the United Nations finally prompted Facebook to begin to change its moderation of content in Myanmar.

[104] *See Social Media Stats Myanmar*, Statcounter, https://gs.statcounter.com/social-media-stats/all/myanmar [https://perma.cc/8M5H-7LZ5] (last visited Oct. 7, 2020).

In 2011, Myanmar began to transition out of nearly five decades of military rule, and in 2013, the new government began the process of liberalizing the telecommunications sector.[105] The impact was quickly apparent in the lives of millions of Burmese. Prior to 2011, a SIM card cost the equivalent of US$1,500 and could only be secured by those connected to the military.[106] By 2013, the cost plummeted to $1.50 and mobile phone shops became ubiquitous.[107]

For a population with no digital literacy, Myanmar's first introduction to the internet came through Facebook, pre-loaded onto their new smartphones by local mobile phone providers.[108] This was not a population that had ever known email, Google, or online news sites; they moved directly from the pre-digital age onto Facebook. As is now commonly noted, in Myanmar "Facebook is the internet."[109] In a population of 54 million people, there are 20 million active Facebook users.[110] This is the same number of people who have internet access.[111] As one research study of users in the capital, Yangon, explained, "users made extensive use of the Facebook search bar similarly to how users in Western nations use search engines."[112] Indeed, most Myanmar users have no concept of looking outside Facebook for independent sources of news or information.[113] This was the backdrop against which genocide began to unfold against the country's minority Rohingya population.

Soon after Facebook was launched in Myanmar, officials from Myanmar's military worked to establish fake Facebook accounts, posing as celebrities and other popular figures. After attracting millions of followers, the officials then spread propaganda through these accounts, targeting the Rohingya.[114] The Rohingya have been subject to persecution in Myanmar for decades, with basic rights, including access to citizenship, progressively stripped

[105] Daniel Thomas & Gwen Robinson, *Myanmar Opens up New Telecoms Frontier,* FIN. TIMES (Apr. 4, 2013), https://www.ft.com/content/118ef1a8-9d17-11e2-88e9-00144feabdc0#axzz48LI4Q5hf [https://perma.cc/QS63-2RXC].

[106] Ei Myat Noe Khin, *What is Myanmar's Digital Rights Situation?*, MEDIUM (Mar. 26, 2018), https://medium.com/@malkhin/what-is-myanmars-digital-rights-situation-de4401c54a63 [https://perma.cc/NP35-7CA2].

[107] Lorian Leong, *Mobile Myanmar: The Development of a Mobile App Culture in Yangon*, 5 MOBILE MEDIA & COMM. 139, 140 (2017).

[108] BSR, HUMAN RIGHTS IMPACT ASSESSMENT: FACEBOOK IN MYANMAR 1, 12 (2018), https://fbnewsroomus.files.wordpress.com/2018/11/bsr-facebook-myanmar-hria_final.pdf [https://perma.cc/7ER8-7Q6U].

[109] *Id.*

[110] Myan. FFM Rep., *supra* note 41 ¶¶ 1342–49.

[111] BSR, *supra* note 108, at 13 ("There are equal numbers of internet users and Facebook users in Myanmar.").

[112] Leong, *supra* note 107, at 147.

[113] BSR, *supra* note 108, at 12–13.

[114] Paul Mozur, *A Genocide Incited on Facebook, With Posts from Myanmar's Military Said to Behind Facebook Campaign That Fueled Genocide*, N.Y. TIMES (Oct. 15, 2018), https://www.nytimes.com/2018/10/15/technology/myanmar-facebook-genocide.html [https://perma.cc/SF7H-B26L]. *See also* Paul Mozur (@paulmozur), TWITTER (Oct. 15, 2018, 8:21 AM), https://twitter.com/paulmozur/status/1051865711155408896 [https://perma.cc/XH3Q-BRWM].

away.[115] Even so, the viral propaganda machine that began to operate against the Rohingya in 2014 was unprecedented. One representative post pictured an advertisement for Rohingya-style food, along with a caption using a derogatory term for the Rohingya, *kalar*: "We must fight them the way Hitler did the Jews, damn *kalars*!"[116]

However, Facebook took no notice, and civil society activists had a hard time raising the alarm. First, the online reporting system available in other locations did not function on the Messenger service through which most people in Myanmar access Facebook, because Facebook's reporting tool used Unicode, the international standard for text encoding, whereas Burmese script uses a typeface called Zawgyi.[117] Additionally, Facebook had entered the Burmese market without translating its Community Guidelines into Burmese[118] and there were no Burmese-speaking Facebook staff.[119]

On occasion, Facebook's policy team would send a delegation to Myanmar. Myanmar-based civil society users routinely flagged inciting content to the platform, but change was slow to come. For example, civil society activists tried to alert Facebook to specific calls to violence against the Rohingya spread through Facebook beginning on September 6, 2017, with a target date for violence of September 11. While Facebook did not detect the messages, activists did and worked through the English-speaking head of a local digital rights organization to alert senior leadership at Facebook.[120] At 6:41AM on September 10, three days after the messages had gone viral, Facebook asked civil society activists in Myanmar for help in identifying the source of the messages. Three violent incidents followed, but seven months later, activists reported that they never heard anything back from Facebook about whether it had found the source of the messages or what action it had taken to stop their further spread.[121]

[115] *See, e.g.*, Kate Cronin-Furman, *The World Knew Ahead of Time the Rohingya Were Facing Genocide,* Foreign Pol'y (Sept. 19, 2017, 4:06 PM), https://foreignpolicy.com/2017/09/19/the-world-knew-ahead-of-time-the-rohingya-were-facing-genocide [https://perma.cc/3USK-NAGH].

[116] Stecklow, *supra* note 42.

[117] Nick LaGrow, *Integrating Autoconversion: Facebook's Path from Zawgyi to Unicode,* Facebook Engineering (Sept. 26, 2019), https://engineering.fb.com/android/unicode-font-converter [https://perma.cc/WM4Z-879G] (explaining how Facebook introduced font converters between Zawgyi and Unicode in September 2019).

[118] Steve Stecklow, *Facebook Removes Burmese Translation Feature after Reuters Report*, Reuters (Sept. 6, 2018, 12:08 PM), https://www.reuters.com/article/us-facebook-myanmar-hate-speech/facebook-removes-burmese-translation-feature-after-reuters-report-idUSKCN1LM200 [https://perma.cc/M4ME-RHMC] (noting that Facebook posted its Community Guidelines in Burmese for the first time in April 2018).

[119] Letter from Mido Phandeeyar, Burma Monitor, Ctr. for Social Integrity, Equality Myanmar, Myanmar Human Rights Educator Network, to Mark Zuckerberg, Chief Exec. Officer, Facebook (Apr. 5, 2018) (available at https://assets.documentcloud.org/documents/4432469/Myanmar-Open-Letter-to-Mark-Zuckerberg.pdf [https://perma.cc/2VVZ-LCS8]). A Facebook spokesperson stated, in a 2018 interview with Wired magazine, that Facebook in 2013 did have one Burmese speaker, located in Ireland. Timothy Mclaughlin, *How Facebook's Rise Fueled Chaos and Confusion in Myanmar,* Wired (July 6, 2010, 7:00 AM), https://www.wired.com/story/how-facebooks-rise-fueled-chaos-and-confusion-in-myanmar [https://perma.cc/WFD9-5ESL].

[120] Telephone Interview with Digital Rights Activist in Myanmar (Aug. 29, 2018).

[121] *Supra* note 119.

In late 2017, the U.N. High Commissioner for Human Rights said that the violence against the Rohingya was a "textbook example" of ethnic cleansing.[122] It was around this time that Facebook started to come under increasing scrutiny in Western media outlets for its role in allowing fake accounts to proliferate in the context of the 2016 U.S. presidential election. Advocates seized this moment, when Western journalists were approaching Facebook with a more critical eye, to draw attention to the problems they had long experienced with Facebook in Myanmar. As the spotlight on Facebook grew, mainstream media outlets began investing significant resources in investigative reporting to document the way in which Myanmar's military had used Facebook to enable the flow of hate speech and incitement against the Rohingya.[123] In an April 2018 interview, CEO Mark Zuckerberg said Facebook had responded by hiring "dozens" of Burmese language content reviewers to monitor hate speech.[124]

In August 2018, the U.N. Independent International Fact-Finding Mission on Myanmar accused the Myanmar military of genocide and war crimes, and described Facebook as a "useful instrument for those seeking to spread hate."[125] In response, Facebook removed eighteen accounts and fifty-two pages associated with the Myanmar military, including the page of the armed forces' commander-in-chief, stating that it wants "to prevent them from using our service to further inflame ethnic and religious tensions."[126] Facebook continues to fail to adequately monitor fake accounts, exacerbated in 2020 by COVID-19 and the laying off of many content moderators.[127]

[122] Michael Safi, *Myanmar Treatment of Rohingya Looks Like "Textbook Ethnic Cleansing," Says UN,* GUARDIAN (Sept. 11, 2017, 6:16 PM), https://www.theguardian.com/world/2017/sep/11/un-myanmars-treatment-of-rohingya-textbook-example-of-ethnic-cleansing [https://perma.cc/BU3U-CCP9].

[123] For a particularly excellent example, *see* Stecklow, *supra* note 42.

[124] Kevin Roose and Paul Mozur, *Zuckerberg Was Called Out Over Myanmar Violence. Here's His Apology*, N.Y. TIMES (Apr. 9, 2018), https://www.nytimes.com/2018/04/09/business/facebook-myanmar-zuckerberg.html [https://perma.cc/99Q5-9Q35].

[125] *See* Myan. FFM Rep., *supra* note 41.

[126] Hannah Ellis-Peterson, *Facebook Removes Accounts Associated with Myanmar Militar*y, GUARDIAN (Aug. 27, 2018, 7:14 AM), https://www.theguardian.com/technology/2018/aug/27/facebook-removes-accounts-myanmar-military-un-report-genocide-rohingya [https://perma.cc/7JRM-CP9E].

[127] *See* Fanny Potkin & Poppy McPherson, *"Spreading Like Wildfire": Facebook Fights Hate Speech, Misinformation Before Myanmar Poll,* REUTERS (Nov. 6, 2020, 6:19 AM), https://www.reuters.com/article/us-myanmar-election-facebook/spreading-like-wildfire-facebook-fights-hate-speech-misinformation-before-myanmar-poll-idUKKBN27L300 [https://perma.cc/E6LG-K8T4] (describing ongoing use of fake accounts in advance of Myanmar's November 2020 election, despite increased efforts by Facebook to tackle the problem); Cat Zakrzewski, *The Technology 202: Posts Promoting Unproven Drugs Highlight Limits of Social Networks' Coronavirus Fight*, WASH. POST (Apr. 27, 2020, 9:11 AM), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/04/27/the-technology-202-posts-promoting-drugs-highlight-limits-of-social-networks-coronavirus-fight/5ea59e8388e0fa3dea9c2b36 [https://perma.cc/4SXY-AZX2] (describing the pandemic's impact on Facebook's human moderators as well as the new policy to attempt to prevent price gouging and fake sales related to the pandemic on the platform). *See also* Yair Rosenberg, Opinion, *Fact-Checking Trump Won't Fix Twitter's Misinformation Problem*, WASH. POST (May 28, 2020, 3:44 PM), https://www.washingtonpost.com/outlook/2020/05/28/twitter-fact-check-trump-voting [https://perma.cc/NK2R-655M] (arguing that the current approach to COVID-19 posting, state-run propaganda, and similar pressing topics should be

Facebook completely dominates the social media market in Myanmar, making it vitally important to Burmese people. But the lack of investment Facebook made in a content moderation system in Myanmar, followed by its unresponsiveness when local civil society groups tried to alert the company to the way state officials were using its platform to fuel violence, suggests that Burmese users are a low priority for Facebook. Without putting necessary resources into a functioning enforcement system, Myanmar provides an early example of the state playing the role of a user and abuser, rather than a regulator, of social media.

2. *India*

Looking to Facebook's launch in India provides a useful counterpoint to the platform's launch in Myanmar. Unlike in Myanmar, Facebook invested enormous resources in the Indian market, opening an office in the country and developing a close relationship with the Indian government. This was not, however, enough to ensure that Indian users could have their concerns about content on the platform addressed. A drop-down menu designed from a Eurocentric perspective failed to include any option for reporting casteist content. Once again, it was advocacy work by civil society groups, not employees of Facebook, that drew attention to this gap.

India, with almost 1.3 billion people, is the most populous country in which Facebook operates and constitutes Facebook's second-largest user base, following the United States.[128] In the wake of President Modi's 2014 election victory, Facebook's Chief Operating Officer Sheryl Sandberg used her visit to India to promote the success of Facebook as a tool for political organizing.[129] Facebook staff had worked with Modi's campaign to help him gain a strong online following.[130] Sandberg highlighted the benefits of using Facebook for electoral purposes: "A perfect example is Narendra Modi, who has extensively used Facebook in his electoral campaign," she ex-

addressed through an automated system, not the current case-by-case system which is less precise and uniform).

[128] Mail Today Bureau, *India is a Big Bet on Facebook Timeline: Sheryl Sandberg,* BUS. TODAY (July 3, 2014, 12:57 PM), https://www.businesstoday.in/technology/news/modi-perfect-example-of-politicians-using-facebook-sheryl-sandberg/story/207817.html [https://perma.cc/EB35-LPH2].

[129] *Id.*

[130] Vindu Goel & Sheera Frenkel, *In India Election, False Posts and Hate Speech Flummox Facebook,* N.Y. TIMES (Apr. 1, 2019), https://www.nytimes.com/2019/04/01/technology/india-elections-facebook.html [https://perma.cc/SSM3-6MN3] ("Facebook worked closely with Mr. Modi's campaign to target ads and rally fans — part of a global effort to spread the use of the platform in politics.").

plained.[131] By 2017, India had overtaken the United States as Facebook's largest market with over 241 million active monthly users.[132]

Facebook opened its first office outside the United States and Europe in Hyderabad, India.[133] In addition to visits to India by Sandberg and Facebook CEO Mark Zuckerberg, President Modi had also made high-profile visits to Silicon Valley, participating in Facebook livestreamed "town hall" meetings with Zuckerberg.[134] As a result, the Indian government enjoys ready access to Facebook staff, and is sufficiently familiar with its reporting structures that it regularly lodges requests with Facebook for the removal of content.[135]

However, these close connections have not shielded Facebook from claims of a lack of cultural competence in its moderation of content in India. In a country of twenty-two official languages in addition to English, Facebook is currently only able to moderate in ten of them, leaving content in a dozen other languages unmoderated.[136] Moreover, the Community Standards to

[131] Mail Today Bureau, *supra* note 128. Sandberg's visit was also part of the build-up to Facebook's 2015 launch of Free Basics in India. Free Basics is an application Facebook developed as a low-cost way to connect people in developing countries to the internet. *Free Basics Platform*, Facebook, https://info.internet.org/en/story/platform [https://perma.cc/G37M-6DG3] (last visited Sept. 16, 2019). Partnering with local telecommunications providers, Free Basics gave mobile phone users a low-data, limited version of internet access for free. The Free Basics package included Facebook and a handful of other websites, with the possibility of additional websites being added if they conformed to the technical requirements Facebook had established. Nanjala Nyabola, *Facebook's Free Basics Is an African Dictator's Dream,* Foreign Pol'y (Oct. 27, 2016), https://foreignpolicy.com/2016/10/27/facebooks-plan-to-wire-africa-is-a-dictators-dream-come-true-free-basics-internet [https://perma.cc/2P4Q-EX7U] (explaining Free Basics and criticizing it in the African context). Ultimately, India's telecommunications regulator banned Free Basics on the ground that it violated India's net neutrality requirements by offering only a limited number of websites for free. Rahul Bhatia, *The Inside Story of Facebook's Biggest Setback,* Guardian (May 12, 2016), https://www.theguardian.com/technology/2016/may/12/facebook-free-basics-india-zuckerberg [https://perma.cc/7PGX-VAHT].

[132] Simon Kemp, *India Overtakes the USA to Become Facebook's #1 Country,* Next Web (July 13, 2017), https://thenextweb.com/contributors/2017/07/13/india-overtakes-usa-become-facebooks-top-country/#.tnw_5GtTSsJi [https://perma.cc/5FZ6-9D92].

[133] PTI, *Facebook to Launch India Operations in Two Months,* Econ. Times (June 27, 2010), https://economictimes.indiatimes.com/tech/internet/facebook-to-launch-india-operations-in-two-months/articleshow/6098727.cms [https://perma.cc/DQQ9-HJ65].

[134] Narendra Modi, *Townhall Q&A with PM Modi and Mark Zuckerberg at Facebook HQ in San Jose, California,* YouTube (Sept. 27, 2015), https://www.youtube.com/watch?v=1-OFfyf7aoA [https://perma.cc/SS38-X7P9]; Narendra Modi, *Townhall Q&A with PM Modi and Mark Zuckerberg at Facebook HQ in San Jose, California,* Facebook (Sept. 27, 2015), https://www.facebook.com/watch/?v=10156199765675165 [https://perma.cc/Y53K-9955].

[135] In the most recent transparency reporting data, July–December 2018, Facebook reported receiving 17,482 content removal requests from the Indian government. Facebook Transparency, *Country Overview: India*, Facebook, https://govtrequests.facebook.com/content-restrictions/country/IN [https://perma.cc/YMB6-MBBR] (last visited Oct. 4, 2019).

[136] Saritha Rai, *Alarming Lessons from Facebook's Push to Stop Fake News in India*, Bloomberg (May 20, 2019), https://www.bloomberg.com/news/articles/2019-05-20/alarming-lessons-from-facebook-s-push-to-stop-fake-news-in-india [https://perma.cc/J3Y9-PZGB].

guide users about what is acceptable speech on the platform are unavailable in languages in which millions of Indian users post content.[137]

Urdu, for example, is the native language of 50 million people in India.[138] It has its own script, which runs right to left. As of the summer of 2019, Facebook's Urdu page for its Community Standards was right-justified, as though it is written right to left, but the Standards are written in English, not Urdu. With native Urdu speakers unaware of what Facebook's standards are with respect to what speech is prohibited on the platform, and unable to use automated reporting forms, there is no way for individuals to use the user-flagging mechanism that usually alerts Facebook to problematic content that has not been filtered out through commercial content moderation or AI.

The moderation challenges in India are not, however, simply linguistic. A rare study of content on Facebook India recently undertaken by Equality Labs, a Southeast Asian civil society organization, tracked the reporting of 1,000 hate speech posts over a four-month period.[139] The study found that the issue of caste constituted a pervasive category of hate speech.

Discrimination on the basis of caste was officially prohibited by India's post-colonial constitution, written in 1950 by Dr. B.R. Ambedkar, who was himself born into the lowest-ranked group in the caste system, the Dalits.[140] Formerly referred to as "the untouchables," Dalits have spent generations being forced into slave labor in jobs such as the manual cleaning of sewage and, even today, they face pervasive social and economic exclusion.[141]

Equality Labs found repeated examples of posts with photo-shopped images of Dr. Ambedkar, accompanied by anti-Dalit slurs. One image, for instance, superimposes Dr. Ambedkar's face onto someone bungee jumping

[137] Thenmozhi Soundararajan et al., *Facebook India: Towards the Tipping Point of Violence: Caste and Religious Hate Speech*, Equality Labs 14 (2019), https://static1.squarespace.com/static/58347d04bebafbb1e66df84c/t/5d0074f67458550001c56af1/1560311033798/Facebook_India_Report_Equality_Labs.pdf [https://perma.cc/DEC3-N3W5] ("the standards are not available in all Indian languages, despite the fact that millions of users speak, read, and write in these languages . . . community standards pages for Indian languages often present only the headings in a regional language and the rest of the text in English.").

[138] Census India, Abstract of Speakers' Strength of Languages and Mother Tongues (2011), http://www.censusindia.gov.in/2011Census/Language-2011/Statement-1.pdf [https://perma.cc/P57J-BLP4] (50,757,762 people listed Urdu as their mother tongue in the 2011 census).

[139] Soundararajan et al., *supra* note 137, at 23.

[140] Billy Perrigo, *As India's Constitution Turns 70, Opposing Sides Fight to Claim Its Author as One of Their Own*, Time (Jan. 24, 2020), https://time.com/5770511/india-protests-br-ambedkar [https://perma.cc/AM7W-8EMX].

[141] *See id.*; Minaketan Bag & Manjulata Jagadala, *Untouchables Amongst Untouchables: An Anthropocentric Study of Ghasi Dalit Women,* 48 Social Change 222 (2018) https://journals.sagepub.com/doi/full/10.1177/0049085718768909 [https://perma.cc/43RH-EBEE] ("even after 70 years of Independence, casteism and untouchability in different forms are still deeply entrenched in Indian society"). Considered to be a caste-based occupation, Dalit women had to remove excrement from both private as well as public dry pit latrines after which they carried the waste off to the disposal grounds. *See also India: Caste Forced to Clean Human Waste*, Hum. Rts. Watch (Aug. 25, 2014), https://www.hrw.org/news/2014/08/25/india-caste-forced-clean-human-waste [https://perma.cc/863L-CK78].

from a bridge, with the caption "Bhangi jumping".[142] "Bhangi" is a derogatory term, similar to the N-word in the American context.[143] If posted offline, such material would violate India's Prevention of Atrocities Act, which was passed in 1989 to support the inclusion of Dalits into Indian society.[144]

Equality Labs tracked efforts to report this and other instances of casteist hate speech to Facebook. However, they found that the drop down menus allowing users to report offensive content did not include a category for hate based on caste, which made it difficult to communicate what it was about these posts that was offensive.[145] Following advocacy efforts by Equality Labs and other activists in the Indian online space, Facebook updated its hate speech policies in 2018 to include caste as a protected category.[146] While only one step in the broader effort to end caste discrimination in India, it means that for the first time since Facebook launched there in 2006, users could point to a rule against caste discrimination to use in their efforts to remove such content from the online public square.[147] This delay suggests that even in markets where a social media company like Facebook is highly motivated to sustain user engagement, designing a content moderation system attuned to the local culture has been challenging.

[142] Soundararajan et al., *supra* note 137, at 43.

[143] *Id.*

[144] The Scheduled Castes and the Scheduled Tribes (Prevention of Atrocities) Act, 1989, India Code, https://www.slideshare.net/OpenSpace/poaord2014 [https://perma.cc/UDW5-QHLH]; Hum. Rts. Watch, India: Events of 2018 (2019), https://www.hrw.org/world-report/2019/country-chapters/india [https://perma.cc/PC55-JL3J] (last visited Sept. 16, 2019) (prohibiting a range of actions, including public insults, against Dalits).

[145] Soundararajan et al., *supra* note 137, at 20–22.

[146] Since 2018, Facebook posts the groups that it considers having protected status for the purposes of hate speech under its public-facing Community Standards. These standards are frequently updated, but it can be difficult to ascertain when exactly changes are made. Certainly prior to 2018, their protected groups did not include caste according to internal Facebook documents, published by the news site ProPublica, which showed defined protected groups as "race, sex, gender identity, religious affiliation, national origin, ethnicity, sexual orientation and serious disability/disease." Julia Angwin et al., *Facebook's Secret Censorship Rules Protect White Men from Hate Speech but Not Black Children,* ProPublica (June 28, 2017), https://www.propublica.org/article/facebook-hate-speech-censorship-internal-documents-algorithms [https://perma.cc/379U-J9AU]. An August 2018 update on hate speech policies posted to the Facebook website is the first documentation I have found that includes caste as a protected category. *Community Standards*, Facebook, https://www.facebook.com/communitystandards/recentupdates/all_updates/ [https://perma.cc/99QV-GS45] (last visited Sept. 18, 2019) (defining protected groups on the basis of "race, ethnicity, national origin, religious affiliation, caste, sexual orientation, sex, gender, gender identity, and serious disability or disease").

[147] Unfortunately, this gain was, many Dalit activists felt, radically undercut by a campaign Facebook launched in March 2019. Intended to inform Indian users about the resources available to them to report hate speech, the campaign featured a harassing fictional troll named Amol Kamble. Kamble is a common surname among Dalits, which led hundreds of users to report the campaign as hate speech that reinforced existing anti-Dalit stereotypes. In the wake of the blunder, Equality Labs urged Facebook to consult with Indian civil society. "Our absence from the discussion table can only lead to more fumbles like this one when it comes to protecting Facebook India's most vulnerable users." Soundararajan et al., *supra* note 137, at 18.

3. *Sri Lanka*

As in Myanmar, Sri Lanka is a place where social media platforms have been instrumentalized to fuel violence. Against the backdrop of decades of inter-ethnic conflict, Sri Lankan officials were alarmed at the way extremists from the Buddhist majority used Facebook to stoke fear of the Muslim minority. Unlike in India, Sri Lankan officials had no direct line to Facebook. In the face of viral misinformation fueling violence against Muslims, and at a time when Facebook changed its algorithm in Sri Lanka to systematically downgrade information disseminated by independent media, the government ultimately resorted to a "self-help" form of content regulation by shutting down access to Facebook altogether. As anti-Muslim misinformation moved from Facebook to Twitter, it became apparent that bots were generating and amplifying the misinformation. And, as the following shows, civil society activists who sought to raise the alarm with Twitter were as unsuccessful as government officials had been with Facebook.

In 2009, Sri Lanka emerged from a 26-year civil war between the majority ethnic Sinhalese (and mostly Buddhist) community and the minority ethnic Tamil community. President Mahinda Rajapaksa and his brother, Defense Secretary Gotabaya Rajapaksa, oversaw the brutal end to the war, with their government forces responsible for the deaths of tens of thousands of Tamil civilians in the final months of the conflict.[148]

President Rajapaksa was re-elected in 2010. In subsequent years, international human rights organizations have worried about an "authoritarian turn" in the country's governance.[149] Local actors speculated about the Rajapaksa family's tacit support for ultra-nationalist groups responsible for violence targeted against Sri Lanka's minority Muslim community.[150]

In 2015, Sri Lankans voted out President Rajapaksa in favor of Maithripala Sirisena, who campaigned on a platform of government reform and reconciliation between the Sinhalese Buddhist majority and the minority Tamil and Muslim communities.[151] Although Sri Lanka's freedom of the

[148] Int'l Crisis Group, War Crimes in Sri Lanka (2019), https://www.crisisgroup.org/asia/south-asia/sri-lanka/war-crimes-sri-lanka [https://perma.cc/J49V-67CG].

[149] *See, e.g.*, Int'l Crisis Group, Sri Lanka's Authoritarian Turn: The Need for International Action (2013), https://www.crisisgroup.org/asia/south-asia/sri-lanka/sri-lanka-s-authoritarian-turn-need-international-action [https://perma.cc/6ETK-4MMK] (noting a turn to authoritarianism under President Rajapaksa's rule).

[150] Tisaranee Gunsekara, *Gotabhaya Rajapaksa and his Bala Sena,* Colombo Telegraph (Mar. 14, 2013), https://www.colombotelegraph.com/index.php/gotabhaya-rajapaksa-and-his-bala-sena [https://perma.cc/863U-5RAZ] (describing connections between Bodu Bala Sena and the Rajapaksa family); Dharisha Bastians & Gardiner Harris, *Buddhist-Muslim Unrest Boils Over in Sri Lanka,* N.Y. Times (June 16, 2014), https://www.nytimes.com/2014/06/17/world/asia/deadly-religious-violence-erupts-in-sri-lanka.html?module=inline [https://perma.cc/VLD9-7PTR] (connecting Bodu Bala Sena to 2014 violence against Muslims in Sri Lanka).

[151] Editorial, *A Step Forward for Sri Lanka*, N.Y. Times (Jan. 14, 2015), https://www.nytimes.com/2015/01/15/opinion/a-step-forward-for-sri-lanka.html [https://perma.cc/34FS-LKS6]; President Maithripala Sirisena of Sri Lanka, Policy Statement to the Eighth Parliament of Sri Lanka (Sept. 1, 2015)

press rating remains low, Freedom House noted recently that freedom of expression had "dramatically improved" since President Sirisena's election.[152]

Sudarshana Gunawardana was responsible for heading Public Information in the new Sri Lankan government. In October 2017, he met with officials from Facebook, urging them to do a better job at moderating hate speech, which he feared could trigger inter-communal violence in the parts of the country where ethnic tensions were high following decades of civil war.[153] Indeed, a Sri Lankan civil society organization, Centre for Policy Alternatives, had already documented how hate speech on Facebook had contributed to a series of Sinhalese attacks on the Muslim community in 2014. The organization noted that a post with thousands of shares and "likes" that clearly violated the platform's Community Standards against hate speech "openly resides on Facebook because it is in Sinhala, a language that [is] outside existing language competencies of Facebook's automated and human-curated monitoring frameworks."[154] Hate and incitement in Sinhalese continued to rise and positions advertised for Sinhalese content moderators went unfilled.[155] Mr. Gunawardana asked for a point of contact to reach out to in an emergency. Facebook did not provide one. Less than six months later, his fears were realized.[156]

In early 2018, a false rumor spread among the majority Sinhalese community on Facebook, claiming that the police had seized 23,000 sterilization pills from a Muslim pharmacist in the rural town of Ampara. After a video falsely suggesting that a Muslim restaurant owner had put a sterilization pill into the soup served to a Sinhalese customer went viral on Facebook, mob violence against the Muslim community in Ampara ensued. Next, a Sinhalese extremist spread a video on Facebook of a driver being beaten in a case of road rage as "evidence" of a Muslim plot to wipe out the Sinhalese population.[157]

With no other means of reaching Facebook, Mr. Gunawardana, other members of President Sirisena's cabinet, and civil society members used Facebook's online reporting system to ask that the videos be removed.[158]

(available at http://www.president.gov.lk/policy-statement-delivered-by-president-maithripala-sirisena-addressing-the-8th-parliament-of-sri-lanka-on-september-1-2015 [https://perma.cc/7CJU-ND66]).

[152] *Freedom in the World 2019: Sri Lanka,* FREEDOM HOUSE, https://freedomhouse.org/report/freedom-world/2019/sri-lanka [https://perma.cc/QU7U-QP4R] (last visited Sept. 18, 2019).

[153] Amanda Taub & Max Fisher, *Where Countries Are Tinderboxes and Facebook is a Match*, N.Y. TIMES (Apr. 21, 2018), https://www.nytimes.com/2018/04/21/world/asia/facebook-sri-lanka-riots.html [https://perma.cc/L9KK-VE5E].

[154] Shilpa Samaratunge & Sanjana Hattotuwa, *Linking Violence: A Study of Hate Speech on Facebook in Sri Lanka*, CTR. FOR POL'Y ALTERNATIVES 4–5 (2014), https://www.cpalanka.org/wp-content/uploads/2014/09/Hate-Speech-Final.pdf [https://perma.cc/6KEZ-VTZU].

[155] Taub & Fisher, *supra* note 153 (noting that twenty-five open job postings for Sinhalese language moderators remained unfilled from June 2017 until April 2018).

[156] *Id.*

[157] *Id.*

[158] *Id.*

Facebook did not respond, and following further online incitement, angry Sinhalese crowds descended on another rural town, the Kandy suburb of Digana, this time causing the destruction of four mosques, thirty-seven houses, forty-six shops and thirty-five vehicles, as well as the death of a 27-year-old man who was trapped inside a burning house.[159] Fearing further violence, the Sri Lankan government took the drastic step of temporarily blocking access to Facebook altogether.[160] Only then did Facebook respond to Sri Lankan officials, arranging for an in-person meeting in the Sri Lankan capital.[161]

The intercommunal violence took place just months after Facebook had changed the way its newsfeed algorithm operated in Sri Lanka. In January 2018, Mark Zuckerberg had announced that Facebook would adjust its newsfeed algorithm so that users saw posts from their friends and family in their main newsfeed, with posts from businesses and media publications set off in a secondary stream.[162] The change was trialed in six countries, including Sri Lanka.[163] While the change was ultimately not made permanent, at least one Sri Lankan media analyst worried about the role it had played in the 2018 violence. "While this experiment lasted, many of us missed out on the bigger picture, on more credible news . . . it's possible that this experiment inadvertently spread hate views in these six countries."[164]

For the three-day period that Facebook was blocked, sophisticated users in Sri Lanka were able to use a virtual private network (VPN) to access the platform.[165] Most users, though, turned to Twitter for information.[166] It was during the attacks in Digana that the Centre for Policy Alternatives saw a tweet on their Twitter feed with a photo that purported to be of "Muslim people in #kandy #digana waiting to attack innocent Sinhalese."[167] Upon

[159] Mujib Mashal & Dharisha Bastians, *Sri Lanka Declares State of Emergency After Mob Attacks on Muslims*, N.Y. Times (Mar. 6, 2018), https://www.nytimes.com/2018/03/06/world/asia/sri-lanka-anti-muslim-violence.html [https://perma.cc/32P2-ZYFF].

[160] Tim McKay, *Sri Lanka Blocked Facebook This Week for Allegedly Spreading Hate Speech and Violence*, Gizmodo (Mar. 10, 2018), https://gizmodo.com/sri-lanka-blocked-facebook-this-week-for-allegedly-spre-1823676276 [https://perma.cc/E26E-WXNN].

[161] Taub & Fisher, *supra* note 153.

[162] Mark Zuckerberg, *Post on Change to Newsfeed: Less Public Content*, Facebook (Jan. 11, 2018), https://www.facebook.com/zuck/posts/10104413015393571 [https://perma.cc/TDS6-UH2H] (last visited Sept. 18, 2019).

[163] Alex Hern, *Facebook Moving Non-Promoted Posts out of News Feed in Trial,* Guardian (Oct. 23, 2017), https://www.theguardian.com/technology/2017/oct/23/facebook-non-promoted-posts-news-feed-new-trial-publishers [https://perma.cc/3VSQ-3DX7].

[164] Taub & Fisher, *supra* note 153.

[165] Once configured, a VPN enables you to send encrypted data through another computer, thereby masking your own IP address and allowing you to access sites that are blocked from your own computer. *See How to Circumvent Online Censorship*, Electronic Frontier Found., https://ssd.eff.org/en/module/how-circumvent-online-censorship#4 [https://perma.cc/7348-DYGX] (last visited Oct. 18, 2019).

[166] Sanjana Hattotuwa et al., Weaponising 280 Characters: What 200,000 Tweets and 4,000 Bots Tell Us About State of Twitter in Sri Lanka 1, 4 (2018), https://www.cpalanka.org/wp-content/uploads/2018/04/Weaponising-280-characters.pdf [https://perma.cc/4U7E-DVK2].

[167] Cat (@cattomm2), Twitter (Mar. 7, 2018 10:11 AM), https://twitter.com/cattomm2/status/971448347927236611 [https://perma.cc/FRY6-L3DD].

noticing some unusual characteristics of the account that posted the tweet—it had no meaningful bio, a prior record of gibberish or nonsensical tweets, and was posting from Frankfurt, Germany—and finding no corroboration of the claim in the tweet, the Centre used their Twitter account to send out a warning about false information.[168] Soon thereafter, the Centre noticed a "tsunami" of Twitter bots following them.[169] They were not alone. In the second half of March 2018, other Sri Lankan activists and journalists reported being followed by some twenty to forty new bots each day.[170]

The bots had been created in batches, and with Sinhala, Tamil or Muslim names, they showed characteristics similar to the bots that the Centre had previously found to be following the Twitter account of Sinhalese politician Namal Rajapaksa, former President Rajapaksa's son.[171] While the Rajapaksa bots amplified content in support of the Rajapaksa family, the flood of bots created in March 2018 were mostly dormant.

Mainstream media coverage of the role that Russian state-sponsored Twitter bots played in influencing the 2016 U.S. presidential election had shown how fake followers could sway political sentiment.[172] The Centre worried that the Sri Lankan bots "will be weaponised in the future, perhaps to control/shape political discourse." They also acknowledged that without greater resources than Sri Lankan civil society organizations have to devote to investigating the phenomenon, it is hard to move beyond speculation. The Centre reached out to Twitter through every channel it had access to. However, "[n]ot a single tweet or email sent by the authors has generated even a canned response from senior company representatives or official accounts."[173]

4. *South Sudan*

Unlike in Myanmar, India, or Sri Lanka, the government of South Sudan has virtually no capacity to engage in the use or regulation of social media. Moreover, unlike in the previous case studies, only a fraction of the South Sudanese population has internet access. Nonetheless, social media impacts the lives of millions of South Sudanese as material spread online by diaspora

[168] Groundviews (@groundviews), TWITTER (Mar. 8, 2018 6:68 PM), https://twitter.com/groundviews/status/971943162062159872 [https://perma.cc/3L2T-SCQJ] ("Be cautious about news reported . . . on Twitter (+ other social media). Evidence suggests bots (aided by trolls) spreading false information with malicious intent.")

[169] Hattotuwa et al., *supra* note 166, at 4.

[170] *Id.*

[171] Sanjana Hattotuwa & Yudhanjaya Wijeratne, *Namal Rajapaksa, Bots and Trolls: New Contours of Digital Propaganda and Online Discourse in Sri Lanka*, GROUNDVIEWS (Jan. 24, 2018), https://groundviews.org/2018/01/24/namal-rajapaksa-bots-and-trolls-new-contours-of-digital-propaganda-and-online-discourse-in-sri-lanka [https://perma.cc/M873-AMY5].

[172] John Reed, *South-East Asia Sprouts Fake Followers for Prominent Twitter Users*, FIN. TIMES (Apr. 18, 2018), https://www.ft.com/content/e59dba5a-421d-11e8-803a-295c97e6fd0b [https://perma.cc/NZZ5-W5RT].

[173] Hattotuwa et al., *supra* note 166, at 15.

communities is shared widely with local communities via text message and then word of mouth.[174] With virtually no independent media in the young nation, this online material becomes a dominant source of information, and against the backdrop of generations of inter-ethnic conflict, online rumors spread rapidly. However, like hate speech everywhere, it is highly contextualized and hard for content moderators not versed in the local culture to identify. As shown in the following, with no government oversight and a dearth of independent media, the local population has been highly vulnerable to hate speech and misinformation.

South Sudan gained its independence from Sudan in 2011, ending Africa's longest-running civil war.[175] The conflict left over two million dead,[176] and caused millions more to flee, creating a worldwide network of diaspora, with large Southern Sudanese communities forming as far away as the United States, Canada, and Australia.[177]

Despite significant oil reserves, the landlocked nation was born one of the poorest in the world, thanks to generations of war and neglect, first by British colonizers and then by the Islamist Sudanese government based in the north of the country. At independence, South Sudan, which is about the size of Texas, had just forty miles of paved roads,[178] and only two percent of its adult population had completed primary school.[179]

Under the terms of South Sudan's independence, a transitional government, led by President Salva Kiir and Vice-President Riek Machar, was to govern the new nation until its first elections in 2015.[180] Kiir hailed from South Sudan's largest ethnic group, the Dinka, and Machar from the second-largest group, the Nuer. But before any election was held, Kiir accused Machar of plotting to overthrow him.[181] The 2013 dispute quickly spiraled into an ethnic conflict that has since killed an estimated 400,000 people and

[174] Freddie Carver, *Webs of Peace and Conflict: Diasporic Engagement in South Sudan*, 29 Horn Afr. Bull. 22, 22–23 (May 4, 2017).

[175] Mary B. Sheridan & Rebecca Hamilton, *South Sudan Seceding Amid Tensions*, Wash. Post (July 7, 2011), https://www.washingtonpost.com/world/africa/south-sudan-secedes-amid-tensions/2011/07/07/gIQAQ8RT2H_story.html [https://perma.cc/XD7L-SBVA].

[176] Human Rights Watch, Human Rights in Southern Sudan (2006), https://www.hrw.org/legacy/backgrounder/africa/sudan0306/6.htm [https://perma.cc/M2LH-8J9A].

[177] Cedric Barnes et al., Rift Valley Inst., The Role of Transnational Networks and Mobile Citizens in South Sudan's Global Community: A Pilot Study Focused on Melbourne and Juba 1, 5 (2018).

[178] Tell Me More, *South Sudanese Rejoice on Eve of Independence*, WAMU: Nat'l Pub. Radio (July 8, 2011, 12:00 PM), https://www.npr.org/2011/07/08/137701357/south-sudanese-rejoice-on-eve-of-independence [https://perma.cc/8KB5-JNUK].

[179] Daniel Maxwell et al., *Researching Livelihoods and Services Affected by Conflict: Livelihoods, Basic Services and Social Protection in South Sudan*, (Feinstein Int'l Ctr., Working Paper No. 1, July 2012) https://www.odi.org/sites/odi.org.uk/files/odi-assets/publications-opinion-files/7716.pdf [https://perma.cc/4ATA-8T8P].

[180] Transitional Const. S. Sudan, 2011, art. 100.

[181] Isma'il Kushkush, *President Says a Coup Failed in South Sudan*, N.Y. Times (Dec. 16, 2013), https://www.nytimes.com/2013/12/17/world/africa/attempted-coup-in-south-sudan-president-says.html [https://perma.cc/WY5G-XRUW].

led a further 2.5 million to flee to neighboring countries in what is now a South Sudanese civil war.[182]

Independent media never gained a strong foothold in South Sudan, due to both the decades of war and the severe state censorship regime run by the Sudanese government. The fledgling South Sudanese government prioritized responding to (and participating in) the new civil war over developing its regulatory capacity with respect to offline or online communications.[183] When the government finally did create a media authority in 2016, the agency opened without office space, and with limited mobile or internet capacity.[184]

In a country where only an estimated seventeen percent of the population has internet access, it would be easy enough to underestimate the impact of online content on South Sudan.[185] However, there is a vibrant community focused on events in South Sudan on Facebook, WhatsApp, and, to a lesser degree, Twitter.[186] Communications on social media are especially high among diaspora, who post information that is subsequently shared inside South Sudan by text messages sent by the few that have internet access to the many who have mobile phones, and then by word of mouth to those who have neither.[187] The information scarcity on the ground, coupled with the esteem in which members of the diaspora are often held, mean that content that is posted by diaspora on social media is often spread widely and uncritically offline in South Sudan.[188]

As in other locations, hate speech in South Sudan is highly contextualized and challenging for anyone unfamiliar with the culture to understand. Peacelab, a civil society actor based in the United States, worked with South Sudanese people to monitor social media throughout 2017 and found significant amounts of hate speech and ethnically framed false news, something that is especially dangerous given the history and ongoing commission of inter-ethnic violence.[189] For example, South Sudanese people began using

[182] Francesco Checchi et al., Estimates of Crisis-Attributable Mortality in South Sudan, December 2013–April 2018: A Statistical Analysis (2018), https://crises.lshtm.ac.uk/2018/09/26/south-sudan-2 [https://perma.cc/8EAV-EX8E].

[183] United Nations Mission in South Sudan, Report on the Right to Freedom of Opinion and Expression in South Sudan Since the July 2016 Crisis (2018), https://www.ohchr.org/Documents/Countries/SS/UNMISS-OHCHR_Freedom_of_Expression.pdf [https://perma.cc/JQE7-EX8C].

[184] *UNESCO Welcomes the Establishment of South Sudan Media Authority*, United Nations Educational, Scientific and Cultural Organization (Feb. 22, 2016), https://en.unesco.org/news/unesco-welcomes-establishment-south-sudan-media-authority-0 [https://perma.cc/D7VT-2VCN].

[185] *Digital 2019: South Sudan*, Data Reportal (Jan. 2019), https://www.slideshare.net/DataReportal/digital-2019-south-sudan-january-2019-v01 [https://perma.cc/ZYU8-2RHG].

[186] *See id.* at slide 23.

[187] Carver, *supra* note 174.

[188] Barnes et al., *supra* note 177, at 23.

[189] For a series of reports documenting ethnically-framed speech online in the South Sudan content in 2017, *see generally Combatting Hate Speech*, Peacetech Lab, https://www.peacetechlab.org/hate-speech [https://perma.cc/48NW-JKW7] (last visited Sept. 21, 2019).

the term "MTN" to refer to the ethnic Dinka population. MTN is one of the major telecommunications companies in East Africa, and its tagline is "everywhere you go." MTN thus plays on the fact that the Dinka, as the largest ethnic group in Southern Sudan, are perceived to be "everywhere you go."[190] As the 2013 ethnic conflict played out, MTN morphed to also become a term associated with targeted killings of Dinka, especially along the main highway leading out of the capital, Juba.[191] As so-called "MTN killings" were discussed online, ethnic tension between the Dinka and other ethnic groups escalated offline.[192] Peacelab reached out to Facebook, and eventually the most inflammatory content was removed.[193] However, MTN is just one of many examples where, when social media companies fail to hire content moderators with local linguistic and cultural competence, they inevitably end up dependent on the (unpaid) labor of civil society organizations to alert them to content that needs monitoring.

Another example that Peacelab identified that further demonstrates how speech is highly contextualized concerns a post by Albert Tamboura, a member of the South Sudanese diaspora. Tamboura refers to himself as "the prophet of South Sudan" and fills his Facebook page with conspiracy theories and angry posts about his child custody battle.[194] In this instance, Tamboura posted a Facebook live video to "alert" his viewers in South Sudan that "Paul Malong Awan [a member of Kiir's cabinet] is planning to kill Equatorians in Juba tonight."[195] The information was not true, but as South Sudanese peace activist Michael Tut explains, when the population in Juba started receiving the message, "they believed it. They didn't know this was false information."[196]

There is nothing in the Tamboura video that would come close to violating the existing Community Standards on Facebook. It is inaccurate, but that alone is not grounds for removal.[197] If his video had warned viewers about an imminent attack by a government official in an offline environment with an operating rule of law, such as the United States or the U.K.,

[190] PeaceTech Lab, Social Media and Conflict in South Sudan: A Lexicon of Hate Speech Terms 7 (2016).

[191] *Three Passenger Buses Attacked on Juba-Nimule Highway*, Sudan Trib. (Oct. 11, 2016), https://www.sudantribune.com/spip.php?article60489 [https://perma.cc/548N-6V4D].

[192] Tom Miles, *U.N. Urges South Sudan to Defuse Growing Tensions Between Ethnic Groups*, Reuters (Oct. 25, 2016, 12:33 PM), https://www.reuters.com/article/us-southsudan-un-idUSKCN12P2BO [https://perma.cc/2WQC-LL6W].

[193] Author Interview with Caleb Gichuhi, Oct. 9, 2019 [on file with author].

[194] *See, e.g.*, Albert Raphael Tamboura, Facebook (Oct. 8, 2020, 8:24 AM), https://www.facebook.com/albert.r.tamboura/posts/10158290884388557 [https://perma.cc/DL5C-2BBG].

[195] *See Combatting Hate Speech*, *supra* note 189.

[196] PeaceTech Lab, *Hate Speech in South Sudan: Social Media as a Weapon of War (Full)*, YouTube (May 3, 2017, 6:27 AM), https://youtu.be/Isz-HugvdDc [https://perma.cc/77FL-MX6M] (Michael Tut, speaking about the spread of the rumor).

[197] *Community Standards: Integrity and Authenticity*, Facebook, https://www.facebook.com/communitystandards/integrity_authenticity [https://perma.cc/W6ZX-Q5Z9] (last visited Sept. 21, 2019) ("we don't remove false news from Facebook").

content moderators following Facebook's global Community Standards likely would have dismissed it. However, Tamboura posted the video just days after 300 people had been killed in Juba.[198] Moreover, Paul Malong Awan, the official Tamboura warned of, had been responsible for South Sudanese military operations that killed hundreds of civilians at the outbreak of the 2013 conflict.[199] If Facebook had a set of content moderators looking to apply their global standards with due deference to this local context, they might reasonably have concluded the misinformation had the potential to inspire violence.[200] However, without local content moderation through the platform or government oversight, and with no independent media to fact check the claims, ethnically driven misinformation like this spreads unchecked among local communities, creating an increased risk of violence.

5. *Turkey*

In turning to Turkey, the following shows how dynamic the interactions between different actors can be when it comes to determining what content appears online. When YouTube first launched in Turkey, the Turkish state had no mechanism for online content regulation.[201] However, a major clash between YouTube's global standards and local Turkish law catalyzed the government's engagement. As in other locations, as the state sought to censor, (or, in more neutral terms, "regulate") online material posted on one platform, users and journalists, rerouted to other platforms. The state in turn became adept at using the tools offered by social media—in this case Twitter's mechanism for geo-blocking content in a given location—to continue to advance its repressive goals.

Over the past twenty years, one political figure has loomed large over Turkey: President Recep Tayyip Erdoğan. Erdoğan founded the current ruling party, Adalet ve Kalkanma Partisi ("AKP") in 2001. In 2003, he became prime minister under what was then a parliamentary governance system.[202] Social media came to Turkey during the first of Erdoğan's three

[198] Jason Burke, *More than 300 Dead as South Sudan Capital is Rocked by Violence*, GUARDIAN (July 11, 2016, 4:57 AM), https://www.theguardian.com/world/2016/jul/10/south-sudan-capital-juba-violence-salva-kiir [https://perma.cc/4ZCJ-KZZT].

[199] He was subsequently placed under sanction by the U.N. Security Council for his role in the 2013 killings. *See Narrative Summary: Paul Malong Awan*, UNITED NATIONS SECURITY COUNCIL, https://www.un.org/securitycouncil/content/paul-malong-awan [https://perma.cc/9WM3-P766] (last visited Sept. 10, 2019).

[200] *See Combatting Hate Speech*, *supra* note 189.

[201] *Attacks on the Press 2007: Turkey*, COMM. TO PROTECT JOURNALISTS (Feb. 5, 2008, 10:20 AM), https://cpj.org/2008/02/attacks-on-the-press-2007-turkey-1.php [https://perma.cc/B56Q-TMNP] ("While the Internet is mostly unregulated, in recent years the courts have ordered the censorship or banning of Web sites deemed in breach of Turkish law.").

[202] *Recep Tayyip Erdogan Fast Facts*, CNN, https://www.cnn.com/2015/11/26/middleeast/recep-tayyip-erdogan-fast-facts/index.html [https://perma.cc/8RFU-4KNB] (last visited Sept. 21, 2019).

terms as prime minister. Today, with a population of eighty-three million, more than sixty percent of Turkish people are active social media users.[203]

The Turkish government's relationship with social media faced an early challenge in 2007 after Greek nationals posted videos on YouTube that insulted Turkey's founding father, Mustafa Kemal Atatürk. Insulting Atatürk is a crime under Turkish law and Turkish courts responded by ordering Turkish telecommunications companies to block access to YouTube.[204] In ensuing negotiations, Google, which owns YouTube, agreed to geo-block the videos inside Turkey.[205] Initially this led Turkey to lift the ban, but then the Turkish government insisted that videos insulting Atatürk be removed worldwide; Google refused, and in 2008, Turkey reinstated the ban on the site.[206]

Until that point, the internet had been largely unregulated in Turkey.[207] However, in the wake of the high-profile dispute, the Turkish parliament passed Law No. 5651, colloquially known as the Internet Law.[208] The law enabled both the judiciary and the executive to block websites under the vague standard of "sufficient suspicion" that an enumerated crime, including insulting Atatürk, had been committed on the website.[209] It thus became legal under Turkish law to block virtually any content through a process that would not withstand scrutiny under the international human rights standards by which Turkey was bound to abide.[210]

[203] Digital 2019: Turkey, https://www.slideshare.net/DataReportal/digital-2019-turkey-january-2019-v01 [https://perma.cc/MQ55-XLH4] (last visited Sept. 21, 2019) (2019 statistics estimate 52 million Turks are active social media users).

[204] Tom Zeller Jr., *YouTube Banned in Turkey After Insults to Ataturk*, N.Y. Times: Lede (Mar. 7, 2007, 9:59 AM), https://thelede.blogs.nytimes.com/2007/03/07/youtube-banned-in-turkey-after-insults-to-ataturk [https://perma.cc/2JVF-LN3Z].

[205] Jeffrey Rosen, *Google's Gatekeepers*, N.Y. Times (Nov. 28, 2008), https://www.nytimes.com/2008/11/30/magazine/30google-t.html [https://perma.cc/U6JK-YHXC].

[206] *Id.*

[207] *Attacks on the Press 2007: Turkey*, Comm. to Protect Journalists (Feb. 5, 2008, 10:20 AM), https://cpj.org/2008/02/attacks-on-the-press-2007-turkey-1.php [https://perma.cc/B56Q-TMNP] ("While the Internet is mostly unregulated, in recent years the courts have ordered the censorship or banning of Web sites deemed in breach of Turkish law.").

[208] 5651 sayili Internet Ortaminda Yapilan Yayinlarin Düzenlenmesi ve Bu Yaynlar Yoluyla Islenen Suçlarla Mücadele Edilmesi Hakkinda Kanun [Law No. 5651 on Regulating Broadcasts Made on the Internet and Combating Crimes Committed Through Such Broadcasts], (2007) (Turk.).

[209] Yaman Akdeniz et al., Report of the OSCE Representative on Freedom of the Media on Turkey and Internet Censorship, Org. for Sec. & Coop. Eur. 1, 12 (Jan. 11, 2010), https://www.osce.org/fom/41091?download=true [https://perma.cc/ZQB2-DPVD] (the OSCE found that in 2008-2009, eighteen percent of website blocks were pursuant to a court order, with the remaining eighty-two percent the result of an administrative order); *see also* Yildirim v. Turkey, 2012-VI Eur. Ct. H.R. 505 (the European Court of Human Rights concluding that the Turkish government had violated the petitioner's right to freedom of expression in its implementation of Law 5651).

[210] *See* General Comment No. 34, Article 19, Freedoms of Opinion and Expression, Human Rights Committee, U.N. Doc. CCPR/C/GC/34 (Sept. 12, 2011), ¶¶ 34–35 (stating that in order to meet the standards of Article 19, a restriction on expression must be necessary and the State "must demonstrate in specific fashion the precise nature of the threat." ¶ 36. A publisher of commentary critical of government "can never be considered to be a necessary restriction of freedom of expression." ¶ 42. Additionally, the restriction must "not be overbroad," and the State "must demonstrate in specific and individualized fashion the precise nature of the threat, and the necessity and proportionality of the specific action

As repression grew, Turkish journalists began turning to Twitter to disseminate news.[211] Subsequently, in 2013, when citizens took to the streets in protest against the AKP, Twitter suddenly became a lightning rod for the government. With traditional media stifled domestically, protestors used Twitter to spread news of the protests both at home and abroad.[212] The events led Erdoğan to characterize social media as "the worst menace to society."[213]

Erdoğan's perspective hardened in the build-up to Turkey's 2014 presidential elections. The elections were the first direct vote for a president, following AKP-backed constitutional changes designed to move Turkey away from a parliamentary system. Erdoğan, then Prime Minister, was the leading presidential candidate. As campaigning peaked in March 2014, his office blocked Twitter.[214]

"Twitter, mwitter!" Erdoğan told a campaign rally. "They say, 'Sir, the international community can say this, can say that.' I don't care at all. Everyone will see how powerful the state of the Republic of Turkey is."[215] Twitter responded by sending out instructions for how Turkish users could tweet via text message.[216] And Turkish citizens, long-practiced at skirting press

taken." ¶ 22.); *See generally* AKDENIZ, *supra* note 203 (discussing ways in which the Internet Law is incompatible with the European Convention on Human Rights, to which Turkey is a party).

[211] Robert Mahoney, *Discarding Reform, Turkey Uses the Law to Repress*, COMM. TO PROTECT JOURNALISTS (2011), https://cpj.org/2012/02/attacks-on-the-press-in-2011-turkeys-legal-problem.php [https://perma.cc/B56Q-TMNP]. Meanwhile, the Turkish government increased its repression of traditional media, and by 2012 Turkey held the ignominious title of jailing the most journalists in the world. *See* Dexter Filkins, *Turkey's Jailed Journalists,* NEW YORKER (Mar. 8, 2012), https://www.newyorker.com/news/daily-comment/turkeys-jailed-journalists [https://perma.cc/3FKJ-FXD4] (jailing 49 journalists in 2012).

[212] Olga Khazan, *These Charts Show How Crucial Twitter is for the Turkey Protesters,* ATLANTIC (June 12, 2013), https://www.theatlantic.com/international/archive/2013/06/these-charts-show-how-crucial-twitter-is-for-the-turkey-protesters/276798 [https://perma.cc/8BBZ-YE]. As the government sought to disperse protestors with tear gas, the country's major television network, CNN-Turk, chose to run a three-part documentary on penguins. Kerem Oktem, *Why Turkey's Mainstream Media Chose to Show Penguins Rather Than Protests*, GUARDIAN (June 9, 2013), https://www.theguardian.com/commentisfree/2013/jun/09/turkey-mainstream-media-penguins-protests [https://perma.cc/M8AT-V8TC]. Tech-savvy protestors quickly turned the penguin into a meme that went viral on Twitter, simultaneously exposing the state of media repression in the country and highlighting the government's violence against the protestors. *See, e.g.*, Zeynep Tufekci (@zeynep), TWITTER (June 7, 2013, 9:28 AM), https://twitter.com/zeynep/status/343041706997993472 [https://perma.cc/7EZZ-8NCS].

[213] Will Oremus, *Turkish Prime Minister Blames Twitter for Unrest, Calls it "The Worst Menace to Society",* SLATE (June 3, 2013, 4:55 PM), https://slate.com/technology/2013/06/turkey-protests-prime-minister-erdogan-blames-twitter-calls-social-media-a-menace-to-society.html [https://perma.cc/T2KM-VWN8].

[214] Laura Rozen, *Twitter Blocked, Turkish Users Say,* AL-MONITOR (Mar. 20, 2014), https://www.al-monitor.com/pulse/originals/2014/03/turkey-twitter-ban-erdogan-wipe-out.html [https://perma.cc/S5S6-6RUG].

[215] Sebnem Arsu, *Turkish Officials Block Twitter in Leak Inquiry,* N.Y. TIMES (Mar. 20, 2014), https://www.nytimes.com/2014/03/21/world/europe/turkish-officials-block-twitter-in-leak-inquiry.html?module=inline [https://perma.cc/2QLT-KLSX].

[216] Twitter Public Policy (@Policy), TWITTER (Mar. 20, 2014, 3:30 PM), https://twitter.com/Policy/status/446775722120458241?ref_src=twsrc%5Etfw%7Ctwcamp%5Etweetembed&ref_url=https%3A%2F%2Fwww.buzzfeednews.com%2Farticle%2Fmiriamberger%2Fturkeys-president-blocks-twitter-but-turkey-keeps-tweeting [https://perma.cc/57LH-TYU4].

censorship, began putting graffiti on public spaces and AKP campaign posters with a domain name server (DNS) number that would allow people to get around the block.[217]

Two weeks later, Turkey's Constitutional Court overturned the ban.[218] However, by the time Twitter had already conceded to start geo-blocking content (Twitter's terminology is Country-Withheld-Content, or "CWC") that allegedly violated Turkish law, a policy they had never previously applied in Turkey.[219]

Twitter's head of global public policy flew to Turkey to meet with government officials.[220] The following month, Twitter blocked a number of popular accounts, following a decision by Turkey's Constitutional Court that the audio recordings containing corruption allegations against Erdoğan that had been disseminated on the accounts had breached Erdoğan's privacy.[221]

Erdoğan won the presidential election and has begun to oversee the largest crackdown on the media in recent Turkish history.[222] Following an attempted coup in July 2016, he instituted a state of emergency and closed over 100 media outlets.[223] Journalists who had originally turned to Twitter as a way of circumventing traditional media censorship soon learned how adept the government had become at closing online communication as well.[224]

One case that attracted high-profile attention was that of Azerbaijani journalist Mahir Zeynalov, who worked for a popular Turkish daily newspa-

[217] Abby Ohlheiser, *How Turkey's Twitter Users are Working Around the Ban,* Atlantic (Mar. 21, 2014), https://www.theatlantic.com/international/archive/2014/03/how-turkeys-twitter-users-are-working-around-ban/359415 [https://perma.cc/NKJ7-XB5Z].

[218] Ceylan Yeginsu, *Turkey Lifts Twitter Ban After Court Calls It Illegal,* N.Y Times (Apr 3, 2014), https://www.nytimes.com/2014/04/04/world/middleeast/turkey-lifts-ban-on-twitter.html [https://perma.cc/MZ2K-ABMV].

[219] Vijaya Gadde, *Challenging the Access Ban in Turkey,* Twitter (Mar. 26, 2014), https://blog.twitter.com/official/en_us/a/2014/challenging-the-access-ban-in-turkey.html [https://perma.cc/WX6S-ZWE5].

[220] Orhan Coskun, *Turkey Accuses Twitter of 'Tax Evasion', Calls for Local Office,* Reuters (Apr. 14, 2014), https://www.reuters.com/article/us-turkey-twitter/turkey-accuses-twitter-of-tax-evasion-calls-for-local-office-idUSBREA3D0TY20140414 [https://perma.cc/3FSP-9DYN].

[221] Seda Sezer, *Turkey Twitter Accounts Appear Blocked After Erdogan Court Action,* Reuters (Apr. 21, 2014), https://www.reuters.com/article/us-turkey-twitter/turkey-twitter-accounts-appear-blocked-after-erdogan-court-action-idUSBREA3J0ET20140420 [https://perma.cc/7VHP-FK78].

[222] *See* Zia Weise, *How Did Things Get So Bad for Turkey's Journalists? The Free Press is Always a Casualty Along the Road to Authoritarianism*, Atlantic (Aug. 23, 2018), https://www.theatlantic.com/international/archive/2018/08/destroying-free-press-erdogan-turkey/568402 [https://perma.cc/P9R7-MU3C].

[223] *Turkey Shutters More Than 100 Media Outlets as Purge Continues*, Comm. to Protect Journalists (July 28, 2016), https://cpj.org/2016/07/turkey-shutters-more-than-100-media-outlets-as-pur.php [https://perma.cc/4S2X-JES4].

[224] Eric Neugeboren, *Erdogan Eyes Social Media: The Last for Turkey's Journalists*, VOA News (July 21, 2020), https://www.voanews.com/press-freedom/erdogan-eyes-social-media-last-refuge-turkeys-journalists [https://perma.cc/MFN5-CQ52].

per.[225] Following Zeynalov's reporting on corruption allegations implicating Erdoğan, Erdoğan filed a criminal complaint alleging Zeynalov had "committed a crime by exceeding the limit of criticism."[226] After the Turkish government put him on a list of foreigners barred from entering Turkey, Zeynalov left the country under police escort in 2014.[227] Nonetheless, he continued to use his Twitter accounts—one in English, the other in Turkish—to report on political developments inside Turkey.[228]

When the 2016 media crackdown began, Zeynalov posted a series of tweets with photos of journalists who had been arrested.[229] The next month, the Turkish government filed a request for scores of Twitter accounts, including Zeynalov's, to be blocked under the Internet Law for "promoting terrorism, encouraging violence and crime, threatening public order and national security."[230] The court approved the government's request and sent the order to Twitter.[231] Twitter refused to block Zeynalov's English-language account, but blocked his Turkish language account inside Turkey under its CWC policy.[232]

Twitter's transparency reports show how content removal requests by the Turkish government have skyrocketed since Twitter first started applying the CWC policy to Turkey in 2014.[233] In the first half of 2014, the Turkish government made 186 removal requests.[234] In the latest available report (January–June 2019), it made 8,993 requests, the second highest number of any country in the world.[235] The Committee to Protect Journalists labeled

[225] Humeyra Pamuk & Dasha Afanasieva, *Turkish Paper Says Journalist Expelled for Criticising Erdogan*, Reuters (Feb. 7, 2014), https://www.reuters.com/article/turkey-media/turkish-paper-says-journalist-expelled-for-criticising-erdogan-idUSL5N0LC0UH20140207 [https://perma.cc/HZD5-U2MV].

[226] Sebnem Arsu & Robert Mackey, *Turkey Deports Journalist for Criticizing Government on Twitter,* N.Y. Times (Feb. 8, 2014), https://thelede.blogs.nytimes.com/2014/02/08/turkey-deports-journalist-for-criticizing-government-on-twitter/?_r=0 [https://perma.cc/JJW5-P9T4].

[227] *Id.*

[228] Sheena Goodyear, *Twitter Blocks Journalist's Turkish Tweets from Being Viewed in His Home Country*, CBC (Sept. 29, 2016), https://www.cbc.ca/news/technology/turkey-twitter-mahir-zeynalov-1.3783815 [https://perma.cc/A9NZ-BKUF].

[229] Hayes Brown, *Turkey is Trying to Get Twitter to Block a Journalist for "Instigating Terrorism",* BuzzFeed News (Sept. 26, 2016), https://www.buzzfeednews.com/article/hayesbrown/turkey-got-twitter-to-block-a-journalist-for-instigating-ter [https://perma.cc/H9BP-KD5K].

[230] *Id.*

[231] *Id.*

[232] Goodyear, *supra* note 228.

[233] Vijaya Gadde, *Challenging the Access Ban in Turkey*, Twitter (Mar. 26, 2014), https://blog.twitter.com/official/en_us/a/2014/challenging-the-access-ban-in-turkey.html [https://perma.cc/D94D-LNST].

[234] *Removal Requests*, Twitter (Jan.–June 2014), https://transparency.twitter.com/en/removal-requests.html#removal-requests-jan-jun-2014 [https://perma.cc/K4AM-KS7Z] (last visited Sept. 21, 2019).

[235] *Removal Requests*, Twitter (Jan.–June 2019), https://transparency.twitter.com/en/removal-requests.html [https://perma.cc/7ELS-A5Q3] (last visited Dec. 30, 2019).

its in-depth study of the topic "How Turkey silences journalists online, one removal request at a time."[236]

Twitter's transparency report for the period in which Zeynalov's account was blocked notes: "Whenever possible under Turkish Law, Twitter filed legal objections in response to court orders involving Turkish journalists and news outlets, arguing that those decisions may be contrary to protections of freedom of expression. None of our objections prevailed."[237]

The Turkish state, more than other states in the case studies above, has adapted to the ways in which users have sought to circumvent state efforts to control the flow of information. Unlike in South Sudan, the Turkish government has been willing and able to take on the role of regulator, but in a way that arguably harms, rather than protects, local users. Efforts by global platforms to be responsive to local rules regarding content regulation, such as Twitter's CWC policy, come from a laudable place. However, the Turkish case study shows how a sufficiently motivated and capable state can use these tools in a way that serves the Turkish state at the expense of most Turkish users.

B. *Understanding the Global Public Square*

The examples presented above, while only capturing a small fraction of activity happening in relation to the regulation of online speech, nonetheless offer a vast expansion of the material upon which the triadic literature has thus far drawn to make sense of the global public square. The following section highlights key dynamics arising from this material to which the existing literature has paid insufficient attention, and which are important for policymakers to absorb as they think through how to improve online speech regulation.

1. *The Local Information Environment*

The local information environment has a huge impact on the effect that content moderation decisions—or non-decisions—have on the local population. The triadic literature has focused on online regulation that plays out in the United States or the E.U. In such locations, populations are not emerging from decades of inter-group violence, they have access to independent media, and the rule of law is generally upheld. When these factors are present in the local information environment, it is easy for their significance to fade into the background. However, as the new case studies show, content

[236] Ahmed Zidan, *How Turkey Silences Journalists Online, One Removal Request at a Time,* Comm. to Protect Journalists (Aug. 13, 2018), https://cpj.org/blog/2018/08/how-turkey-silences-journalists-online-one-removal.php [https://perma.cc/28LU-V6R3].

[237] *Removal Requests*, Twitter (July–Dec. 2016), https://transparency.twitter.com/en/removal-requests.html#removal-requests-jul-dec-2016 [https://perma.cc/227Y-GEM8].

moderations decisions play out differently in local contexts where such factors are absent.

In South Sudan, the dearth of access to independent media makes it near-impossible for users to fact-check false information, and against the backdrop of the ongoing experience of inter-communal violence with no state actor stepping up to enforce the rule of law, the spread of false information can bring about deadly outcomes. In Turkey, where the state is exerting evermore control of the media, online regulation that leads a journalist to be blocked from a platform has greater significance than it would in a country where independent media was vibrant.

Both the existing literature and the content moderation systems used by platforms assume generally shared norms by users within a state, with so-called "dangerous speech" the province of a minority of users.[238] Neither the literature nor the regulatory approach of platforms is equipped to respond to scenarios, commonplace in countries emerging from decades of inter-communal violence, in which such speech is tolerated or supported by the majority of a community and/or the state itself. In Myanmar, decades of state-sanctioned persecution of the Rohingya left relatively few users to flag anti-Rohingya content as offensive. In Sri Lanka, following decades of civil war, fear of violence by those viewed as "other" is so deep that false information is readily acted upon with lethal results.

An appreciation for these realities helps us understand why content moderation rules, developed for an information environment that flows from a stable liberal democracy, may not make sense in other contexts in the global public square.[239] It underscores the reality that at least some global rules will have to be localized differently to achieve the same goal in one locale versus another.

[238] For an explanation of this concept, *see* Susan Benesch, *What is Dangerous Speech?,* DANGEROUS SPEECH PROJECT, https://dangerousspeech.org/about-dangerous-speech [https://perma.cc/7Q2Z-NUEC] (last visited Oct. 4, 2019) ("Dangerous Speech is any form of expression (speech, text, or images) that can increase the risk that its audience will condone or participate in violence against members of another group.").

[239] As this Article is being finalized, President Trump is in the midst of contesting the legitimacy of the 2020 U.S. presidential election results, and is using Twitter as his primary forum for broadcasting his message. In response to this development, Twitter has introduced a new policy that involves applying a "disputed and potentially misleading" label to relevant tweets. Chris Mills Rodrigo, *Twitter Says It Labeled 300,000 Posts Around the Election*, HILL (Nov. 12, 2020, 5:49 PM) https://thehill.com/policy/technology/technology/525770-twitter-says-it-labeled-300000-posts-around-the-election?fbclid=IWAR3pUT02m8vWOTY5965uWSYcsp1OT16ppe3PAajezzi9guvTuzwewEgrIoA [https://perma.cc/HNB7-32F3] (noting that between Nov. 3–7, Twitter applied the label to nearly half of President Trump's election-related tweets). This seems to strengthen the argument that policies originally crafted with a stable liberal democracy in mind must be adapted for different contexts—including when the political context changes in the very country in which the policies were originally formulated.

2. *Plurality of Actors and Roles*

i. Beyond the State as Regulator

The case studies highlight the range of roles that a state can take on, as well as how a state can be entirely absent from the regulation of online speech. By looking primarily at the United States and E.U., the triadic literature typically assumes that the state is at least capable of regulating online content, even if in the case of the United States it seems historically disinclined to do so. However, South Sudan represents an example of a state that simply does not have the capacity to regulate online content, regardless of its willingness to do so. Consequently, when applied to South Sudan, the free speech triangle lacks the "state" prong of the triad entirely. Furthermore, given its focus on the United States and the E.U., the existing literature also assumes that a state has the capacity to secure the attention of a social media platform. That assumption holds in India, for instance, where market dynamics led Facebook to engage closely with the Indian government. However, as was clear from the Sri Lanka case study, the assumption that a state is able to get a social media company to respond to its concerns does not hold true globally. Meanwhile, the role of the military in Myanmar demonstrates that the state may act as a major user of social media.[240] The notion of the state as a user of social media has not traditionally been associated with the Global North, although President Trump's Twitter account has been effective in chipping away at that notion.[241] Had social media platforms and the associated literature on them paid more attention to Myanmar from the outset of Facebook's launch in the country, they could have begun contingency planning much earlier for how to respond to misinformation disseminated by state actors worldwide – including, most recently, by the U.S. president.[242] Overall, the triadic model has tended towards a reductionist vision of the role that the state can play within the social media ecosystem. By drawing in a broader range of case studies, it is easier to appreciate the plurality of roles that states can adopt.

ii. Bots, Civil Society, Traditional Media, Diaspora

In addition to diversifying our understanding of the range of roles undertaken by actors who comprise the free speech triangle, the case studies also highlight the important roles that actors who have received comparatively little attention in the existing scholarship can play. As is increasingly recog-

[240] This observation is not limited to the case studies presented here. *See, e.g.*, SANJA KELLY ET AL., FREEDOM HOUSE, FREEDOM ON THE NET 2017: MANIPULATING SOCIAL MEDIA TO UNDERMINE DEMOCRACY (2017) (discussing "state-led interventions" in the Global South, including in Venezuela, Ecuador, Thailand and The Philippines.)

[241] *See, e.g.*, Knight First Amendment Inst. at Columbia Univ. v. Trump, 302 F. Supp. 3d 541 (S.D.N.Y. 2018); Knight First Amendment Inst. at Columbia Univ. v. Trump, 928 F.3d 226 (2d Cir. 2019) (challenging President Trump's action in blocking followers from his Twitter account as a violation of the First Amendment).

[242] *See supra* at note 240.

nized in the U.S. context, bots are a growing force to contend with in the regulation of online content. While it is too soon to tell the impact of the troll armies developing on Sri Lankan Twitter, the company's response to the concerns about the influx of bots thus far is not encouraging.

The case studies highlight the role of civil society groups as well, especially in locations where social media platforms have the least cultural or linguistic competency. Without any Burmese speakers on payroll or contract when it launched in Myanmar, Facebook was entirely reliant on local civil society actors to alert it to the weaponization of its platform. In India, despite ample contact between Facebook and the Indian government, it was civil society that highlighted the need to include caste as a protected category in the Indian context. Across the Global South, such groups take on systematic and large-scale content moderation work, without any social media platform paying them for their labor. Notably, in some locations, there are no civil society groups with the capacity to take on this role. Further, in other places, the civil society groups with the capacity to fill in the labor gap these companies have created are already part of the elite, leaving the most marginalized groups further oppressed.

The case studies also show the significant role of traditional media in online content regulation conducted by the state or by a platform. In a case such as Turkey, for example, journalists were initially successful at circumventing state censorship by moving onto Twitter, but in time these actions catalyzed the state to become adept at using the legal processes provided by Twitter to achieve its censorship goals regarding material posted online. In Myanmar, actions of traditional media were what catalyzed Facebook to increase its online content regulation. Myanmar civil society groups failed to secure any of the changes they requested from Facebook until there was a mainstream media spotlight on Facebook's role in the Rohingya crisis.

Finally, the South Sudan case study shows the way in which online content focused on a particular locale is not limited to users who are physically in that location; diaspora networks are global, even as the content they focus on can be highly localized, and online content produced at a distance can have a significant impact on local offline activity. This same reality is also apparent in the case study from Turkey, with a journalist expelled from the country still able to have a significant impact on the local online conversation. This, of course, is an observation that applies equally to content focused on the Global North. But, again, it is a dynamic that has not yet been highlighted in the triadic literature. Overall, a combination of bots, civil society actors, traditional media outlets, and diaspora networks can have a significant impact on what material does and does not appear online, yet the triadic model makes it too easy to overlook the roles they play.

3. *Fluidity of Jurisdictional Reach*

Online content regulation is, perhaps inescapably, both local and global. Platforms design rules intended for global implementation given the scale at which they operate. Yet, local actors also find ways to secure rules that are tailored to the local context. Sometimes this plurality exists side by side, but at other times the global and local come into direct conflict. Moreover, the case studies suggest that the resolution of these conflicts is determined on the basis of power, rather than principle.

The existing literature has highlighted the ways in which E.U. law can have extraterritorial effect, in other words, the local rule went global. Yet when Turkey sought to have its local rule prohibiting content offensive to Atatürk implemented worldwide, YouTube refused. Platforms have, however, allowed Turkey to deviate from their global rules. These global rules upholding free expression protect criticism of Atatürk, yet in Turkey platforms use geo-blocking to prohibit such criticism. Unlike that of the E.U., Turkey's local rule did not go global, but it did permit Turkey to deviate from the platform's global rule.

Meanwhile, there are globally applicable rules that arguably do not work for users in South Sudan, for example. If the South Sudanese state had the strength and resources of the Turkish state, then it may be able to secure a localized version of the global standard regarding the threshold for material that incites violence. Yet in the absence of a strong state in South Sudan, no localized options have been explored and South Sudanese users are stuck with the global rule. The ability of a state to opt out of a platform's global rule, or secure a locally contextualized version of the rule, should not turn on the power of a state's government. Without any principled way to deal with the need for flexibility in a pluralized system, the ability of local populations to secure appropriately contextualized modifications to global rules seems to turn on state power.

Overall, the case studies add both breadth and depth to our understanding of how the global public square operates. Empirical observation of cases outside the North American and Euro-centric space not only diversifies our understanding of the actors involved in the social media ecosystem, but also highlights the range of roles that can be undertaken by the actors who are included in the triadic model. The effort underscores the difficulty of trying to capture even a simplified version of this reality by using an actor's status as shorthand for the role they play in this ecosystem; not only are there too many actors to capture, but the possible roles they play cannot be defined solely by their status. A turn to function is in order.

III. Toward Better Governance

The ability to accurately describe reality is a necessary but not a sufficient condition for the development of good governance policies. Faced with a messy and complex world, policymakers, scholars, analysts and advocates all benefit from simplifying models to help structure the way they think through complicated problems.

Accepting the global public square as a site of pluralism provides a pathway to structure our thinking about how content makes it online. Pluralism has always demanded that scholars focus on what different actors actually *do*, rather than relying on the identity or status of an actor as shorthand for the activity they undertake.[243] This methodological orientation is invaluable in the current moment when the global public, and even policymakers, are struggling to gain a basic understanding of how online content regulation operates at scale.

The relevance of this focus on function, in addition to status, is underscored by the examples presented in Part II; the global public square is a place where a range of actors each undertake multiple functions in the process of determining what content makes it online. The free speech triangle proposed by Balkin captures the key actors involved in governing the global public square. However, by drawing on a broader array of examples to strengthen the descriptive account of this space, it is possible to supplement the free speech triangle with a functional approach that can help policymakers identify where to focus their efforts.

Once we allow empirical observation, as necessitated by a pluralist approach, to take the lead, it becomes possible to distill the governance of the global public square into to four key activities: What content makes it online is determined through contestation between a plurality of actors engaged in i) content production, ii) content amplification, iii) rule creation, and iv) enforcement.

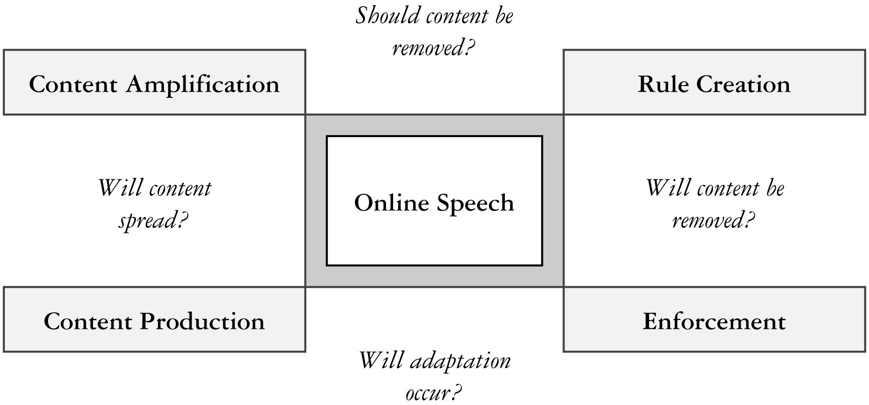
A. A Function-based Approach

A functional approach to the question of what content makes it into the global public square surfaces questions that the existing literature has often overlooked and helps direct attention to the recurring points of tension that we see in this space. The schematic below does not replace the free speech triangle, but rather supplements it. This clarification is important, because the status of different actors remains relevant to both subjective and objective expectations of the actions a given actor can and should take. Normatively and positively, the state holds different rights and is subject to different responsibilities than an individual user, or than a nonstate social media platform. However, as the case studies illustrate, an actor's formal

[243] *See generally* John Griffiths, *What is Legal Pluralism?,* J. Legal Pluralism 1 (1986).

status alone is insufficient as a tool for understanding the way it interacts with online material. By layering function over status, we gain an analytic tool that offers much broader applicability.

FIGURE 2: FUNCTION-BASED MODEL



Below I briefly describe the activities that take place at each point in Figure 2, which represents a function-based approach. The points on the model are delineated by the function undertaken. These functions do not necessarily take place in sequential fashion; content production, content amplification, rule creation, and enforcement take place concurrently and sometimes in relation to the same piece of content.

On each side of the square is a process connecting the adjacent points of the square. Thus, once content production takes place the question is whether that piece of material will spread, perhaps even virally, or will remain unseen by anyone beyond the user who posted it. The answer to this question of "Will content spread?" turns on the adjacent function of content amplification, on the top left-hand corner of the square. Once content is, or is not, amplified, a normative question arises as to whether that content should remain online. The answer to that question turns on the adjacent function of rule creation, at the top right-hand corner of the square. Once a rule regarding the content exists, the question is whether the content will be removed, which in turn is a function of whether the relevant rule is actually enforced, as seen at the bottom right hand corner of the square. And if the rule is enforced, and content is removed, then there is the final question of whether adaptation will occur. Users who wish the now-removed content to be accessible may find novel ways around its removal, perhaps modifying or distorting the original content to generate equivalent material in a new effort at content production, or finding other platforms for dissemination.

The benefit of this approach is its broad applicability; it is as useful in a locale where the state is a strong regulator or utterly absent, where there is a robust civil society or none at all. It accounts for the variety of different actors in play without locking in any preconceptions about which actors undertake which activities. As discussed below, those involved in *content production* include individuals, businesses and governments across the world. Those involved in *content amplification* can include, for example, individuals from South Sudan's diaspora, state-sponsored bots, and YouTube's recommendation algorithm. Those involved in *rule creation* include, for instance, Twitter, the Turkish Presidency, and digital activists based in India. Finally, while technology companies are typically on the front lines of enforcement, some states are also active in this role.

1. *Content Production*

Content production involves the creation of online material—something that billions of individuals, corporations, and states are involved in, on a minute-by-minute basis, around the world. In some cases, the actor who creates the content disseminates it themselves—the cliché of an individual user posting a video of her cat. In other cases, a state-sponsored actor may create the content and amplify it through bots. We saw this in the Myanmar example, and are now seeing it unfold again in relation to COVID-19 and the Black Lives Matter protests.[244] Also falling within this category is content created to modify or distort existing content, something that can be done by individuals, states, private organizations or, increasingly, automated programming.[245]

2. *Content Amplification*

Content amplification can be undertaken by individuals, but this is the part of the system where technology companies and bots play an outsized role. One might think, for instance, of how the Russian-sponsored Internet Research Agency created content for inflammatory Facebook ads to sow discord among American voters, and then amplified engagement with the ads through thousands of bots—fake accounts that liked or shared the content.[246]

To maximize their value to advertisers, social media platforms design their recommendation algorithms to sustain user engagement. While the specifics of these algorithms are proprietary, and thus opaque to the public,

[244] Blackbird.AI, COVID-19 (Coronavirus) Disinformation Report: Volume 2.0 (2020), https://www.blackbird.ai/wp-content/uploads/2020/03/Blackbird.AI-Disinformation-Report-COVID19-Volume-2.pdf [https://perma.cc/W7TQ-WG5N]

[245] With thanks to Tom Dannenbaum for this observation.

[246] Andrew Orlowski, *U.S. Congress Finally Emits All 3,000 Russian 'Troll' Facebook Ads. Let's Take a Look at Some,* Register (May 10, 2018), https://www.theregister.co.uk/2018/05/10/congress_russian_facebook_troll_ads [https://perma.cc/VZ29-FFQ2].

interviews with former employees repeatedly emphasize the way these algorithms push users toward extremist content.[247] We saw this possibility in the Sri Lanka case study, where changes to Facebook's newsfeed algorithm may have helped amplify the anti-Sinhalese content that was promoted through friends and family by systematically minimizing content generated by independent media that could otherwise have served as a fact-based counterpoint to the ethnically driven rumors.[248]

3. *Rule Creation*

Once content is posted online, a normative question arises: Should content be taken offline? The answer to that question involves a regulatory debate that boils down to the question of what rules should apply. The default response is that the technology company's existing rules should apply: Facebook's Community Guidelines, or Twitter's Terms of Service, for example. However, the existing platform guidelines might be disputed and other actors may push for a new rule, as we saw Indian civil society activists do in terms of getting Facebook's Community Guidelines amended to reflect caste discrimination. Likewise, an executive, legislative, or judicial body at the local, state, or regional level may argue that a different rule should apply within their territory, and perhaps even globally. We can think of the example of YouTube in Turkey, where Google agreed to apply the local law prohibiting access to content offensive to Atatürk's memory within Turkey.[249] Alternatively, consider the long-running litigation between Google and France in which the European Court of Justice concluded that while Google needed to remove content from E.U. member states under a right to be forgotten claim under French law, it did not have to do so globally.[250] This *rule creation* portion of the model is legal pluralism in action, as different normative communities come into conflict over what rule should apply. Such regulatory activity may generate a conclusion—either

[247] *See e.g.* Kevin Roose, *The Making of a YouTube Radical,* N.Y. Times (June 8, 2019), https://www.nytimes.com/interactive/2019/06/08/technology/youtube-radical.html [https://perma.cc/BAJ2-RFF4] ("'There's a spectrum on YouTube between the calm section—the Walter Cronkite, Carl Sagan part—and Crazytown, where the extreme stuff is,' said Tristan Harris, a former design ethicist at Google, YouTube's parent company. 'If I'm YouTube and I want you to watch more, I'm always going to steer you toward Crazytown.'"). *See also How YouTube Decides What You Should Watch,* BBC (May 24, 2019), https://www.bbc.co.uk/sounds/play/w3csyvmt [https://perma.cc/TD6P-VYFM].

[248] In another example of amplification of content by a platform, news reports uncovered the role of YouTube's recommendation algorithm in amplifying a home video uploaded by a ten-year-old girl and her friend of them playing in a backyard swimming pool. The video received 400,000 views after being recommended through a network of viewers "seemingly motivated by sexual interest in children." Max Fisher & Amanda Taub, *On YouTube's Digital Playground, an Open Gate for Pedophiles,* N.Y. Times (June 3, 2019), https://www.nytimes.com/2019/06/03/world/americas/youtube-pedophiles.html [https://perma.cc/WKS8-2TXA].

[249] Rosen, *supra* note 207 ("by using a technique called I.P. blocking, [Google] would prevent access to videos that clearly violated Turkish law, but only in Turkey.").

[250] Court of Justice of the European Union Press Release No 112/19, The Operator of a Search Engine is not Required to Carry out a De-referencing on All Versions of its Search Engine (Sept. 24, 2019).

through active debate or by default—over whose rule applies. Yet, consistent with the pluralist nature of the system, that conclusion remains subject to future challenge.

4. *Enforcement*

Regardless of the regulatory conclusion to the normative question, a positivist question arises: Will content be taken offline? This directs us to enforcement, which is most commonly undertaken by the technology company on whose platform the content is posted.

Enforcement activity is happening constantly; sometimes it is directly connected to the application of a rule, but at other times it is not. With the help of AI, platforms can enforce content even before it actually appears online. This, for example, is what happened with a video of the mass shooting perpetrated against New Zealand Muslims in a Christchurch mosque in March 2019; Facebook intercepted 1.2 million attempts to post the video at the point of upload.[251] Alternatively, content can be removed on a timespan ranging from seconds to years after it first appears online, either through AI or by commercial content moderators.

States can also undertake enforcement activity directly by blocking access to the platform entirely, as we saw in the Sri Lanka example.[252] In this scenario, a wide range of content, including that which the state itself has no objection to, is inaccessible for the period of the shutdown.

Unlike technology companies and states, most users cannot undertake direct enforcement action. Hackers or others with similarly sophisticated technical skills may be able to launch the kind of DDoS attacks that temporarily brought down Wikileaks.[253] Other individuals have successfully obtained legal injunctions, though this remains an outlier approach, limited to generally well-resourced and/or supported actors.[254]

There are times when enforcement flows directly from rule creation. When Twitter decided that the Turkish law prohibiting content offensive to the memory of Ataturk should be applied inside Turkey, it enforced removal of previously accessible content. However, rule creation and enforcement are not always so tightly connected. Enforcement actions can involve mistakes where human or algorithmic error leads to the removal content that does not

[251] Guy Rosen, *A Further Update on New Zealand Terrorist Attack,* Facebook (Mar. 20, 2019), https://newsroom.fb.com/news/2019/03/technical-update-on-new-zealand [https://perma.cc/TK7D-8DNJ].

[252] Yet more extreme actions have seen a state block access to the internet entirely, as India recently did in the province Kashmir. *See* Vindu Goel et al., *India Shut Down Kashmir's Internet Access. Now, 'We Cannot Do Anything.'*, N.Y. Times (Aug. 14, 2019), https://www.nytimes.com/2019/08/14/technology/india-kashmir-internet.html [https://perma.cc/E9D4-EL36].

[253] It is a very limited number of users with such a profile, though this may change over time. There is, nonetheless, a degree to which all users can take a limited form of direct enforcement action by "blocking" or "unfollowing" certain accounts.

[254] *See, e.g.*, Case C-18/18, Glawischnig-Piesczek v. Facebook Ireland Ltd, ECLI:EU:C:2019:821 (Oct. 3, 2019).

actually fall afoul of any rule. In the summer of 2017, for example, YouTube removed 200,000 videos of atrocities in Syria that human rights defenders had uploaded in the hope of achieving accountability for war crimes.[255] In prior work, I have explained how the removals did not flow from any rule prohibiting documentation of Syrian atrocities. Rather, it was the result of a new YouTube algorithm that was unable to distinguish between documentation videos uploaded by human rights defenders from terrorist content; the platform was trying to remove the latter following a push by E.U. legislators to apply a rule prohibiting extremist postings.[256] Recently the issue resurfaced as Facebook, and its subsidiary Instagram, deactivated or suspended accounts of numerous journalists, activists, and civilians in Syria, Palestine, and Tunisia due to inaccurate algorithms and COVID-related reduced human oversight. Like YouTube, Facebook could not distinguish between posts exposing human rights violations and those endorsing or engaging in terrorism.[257]

Lastly there is the question of adaptation. How will an actor respond to their experience of the system? Will a technology company that has removed a given piece of content invest in mechanisms that would proactively identify similar content in the future? We have seen an example of the latter in the way that hash technology, originally developed to remove child pornography from the internet, is now being re-purposed by technology companies to improve the rate at which they remove terrorist content.[258] Will an advocacy group that has lost the initial argument over getting a certain type of content removed find a new way to argue for their removal rule to be adopted in the future?[259] Does a user who is banned from a platform have a

[255] Avi Asher-Schapiro & Ban Barkawi, *"Lost Memories": War Crimes Evidence Threatened by AI Moderation*, Reuters (June 19, 2020, 12:43 PM), https://www.reuters.com/article/us-global-socialmedia-rights-trfn/lost-memories-war-crimes-evidence-threatened-by-ai-moderation-idUSKBN23Q2TO [https://perma.cc/U6QX-BUY2].

[256] *See* Rebecca J. Hamilton, *Social Media Platforms in International Criminal Investigations*, 52 Case W. Res. J. Int'l. L. 213 (2020). Mistakes are inevitable given the scale and diversity of content that appears online, but the question of which mistakes are identified and remedied as a matter of priority nonetheless comes down to the intersection of power and interests discussed *infra* Part III.B.3, with technology companies holding disproportionate amounts of power by virtue of their ability to directly remove content.

[257] *See* Olivia Solon, *"Facebook Doesn't Care": Activists Say Accounts Removed Despite Zuckerberg's Free-Speech Stance*, NBC News (June 15, 2020), https://www.nbcnews.com/tech/tech-news/facebook-doesn-t-care-activists-say-accounts-removed-despite-zuckerberg-n1231110 [https://perma.cc/KUY8-LYEB] (describing the deletion of accounts, often without a warning or reason provided in April and May 2020 to include: the thirty-five accounts deleted in Syria, fifty-two accounts deleted in a single day in Palestine, and sixty accounts deleted in Tunisia, the latter of which Facebook admitted some deletions were mistakes).

[258] *See* Julia Franz, *How We Can Use "Digital Fingerprints" to Keep Terrorist Messaging from Spreading Online,* World (Feb. 12, 2017), https://www.pri.org/stories/2017-02-12/how-we-can-use-digital-fingerprints-keep-terrorist-messaging-spreading-online [https://perma.cc/XD58-FAZ9].

[259] *See, e.g.*, Gillespie, *supra* note 67, at 87–97 (describing a successful campaign to get Facebook to change its rules on the posting of breastfeeding photos).

sufficiently large or otherwise influential set of supporters who will re-post the content?[260]

B. *A Roadmap for Policymakers?*

A function-based approach is a useful supplement to the free speech triangle as it challenges our parochial assumptions about which actors undertake which roles and suggests a way to think about the governance of the global public square that has broader applicability. I explore the strengths of this approach in more detail below but it is worth highlighting at the outset a serious concern with the turn to a function-driven approach.

Functionalism in any form risks freezing the status quo. By looking at what various actors are doing in the present moment, and developing policy accordingly, it risks setting existing power dynamics in stone. The function-based approach developed here faces a similar problem. When applied to South Sudan in 2020, for example, the function-based model emphasizes the absence of the state in the regulatory discussion. By building this absence into the model, the function-based approach risks writing the state out of the regulatory space, thereby limiting the conversation about how the state could—and should—step into a regulatory role in the future.

A better approach is to view the value of the function-based model in terms of its ability to highlight relationships and dynamics that are ripe for transformation. Recast in this way, a function-based model provides a roadmap for those seeking to address concerns with the way the global public square is currently governed. In the following section, I highlight some of the relationships and dynamics to which the function-based model draws our attention.

1. *States as Content Producers*

Moving to a function-based approach focuses attention on the possibility of states being involved in content production, both through the creation of novel content and the modification and/or distortion of existing content. This observation seems obvious in the aftermath of Russian interference in the 2016 U.S. presidential election.[261] However, such recognition is disturbingly recent. "We were too slow to spot and respond to Russian interference," Zuckerberg told Congress in April 2018.[262]

[260] This is what happened when YouTube removed Alex Jones's channel. *See* Alex Kaplan, *YouTube is Allowing Multiple Accounts to Circumvent Its Ban Against Alex Jones' Infowars,* MEDIA MATTERS (June 20, 2019), https://www.mediamatters.org/alex-jones/youtube-allowing-multiple-accounts-circumvent-its-ban-against-alex-jones-infowars [https://perma.cc/AZ9D-UK76].

[261] *See* MUELLER, *supra* note 22.

[262] Mark Zuckerberg, Chief Exec. Officer, Facebook, Testimony to the U.S. Senate Committees on the Judiciary and on Commerce, Science and Transportation (Apr. 10, 2018) (available at https://www.judiciary.senate.gov/imo/media/doc/04-10-18%20Zuckerberg%20Testimony.pdf [https://perma.cc/4ZQV-BYJ7]).

Further recognition of this problem can be seen in platforms' current initiatives aimed at labeling and banning some state-run media.[263] However, both internal and external critics are denouncing Facebook's refusal to fact-check or label the U.S. president's incendiary posts when other platforms, including Twitter and Snapchat, have.[264] The function-based model helps policymakers focus on the need to prevent and mitigate damage that states can do through content production—whether by influencing American voters, inciting violence against the Rohingya, spreading misinformation about COVID-19, or the many other scenarios that are likely to arise.

2. *Technology Companies as Content Amplifiers*

The function-based model also highlights the role of technology companies in content amplification. While technology companies do not create content, they have an important role in its dissemination. From Facebook's newsfeed algorithm to YouTube's recommendation algorithm, technology companies' design choices are driven by the goal of keeping users on their platforms.[265]

By highlighting the role of technology companies in content amplification, the function-based model reveals the limits of asking companies to play an enforcement role simply by removing prohibited content. It raises the possibility of a technology company playing an enforcement role, say, by removing a piece of content that incites violence against a Sri Lankan Muslim community, while its algorithm simultaneously recommends similar content to its users in the Sri Lankan Sinhalese community. Thus, when regulators think about how to keep prohibited content offline, they need to think beyond the "whack-a-mole" approach of content removal and also engage with the structural changes that technology companies may need to make to their recommendation algorithms.[266]

[263] Hadas Gold, *Facebook Will Start Labeling Pages and Posts from State-controlled Media*, CNN Bus. (June 5, 2020), https://www.cnn.com/2020/06/04/media/facebook-state-media-label/index.html [https://perma.cc/YLK6-FEF7].

[264] *See* Nitasha Tiku, *Scientists Funded by Zuckerberg Sent Him a Letter Calling Facebook's Practices 'Antithetical' to His Philanthropic Mission*, Wash. Post (June 6, 2020), https://www.washingtonpost.com/technology/2020/06/06/chan-zuckerberg-initiative-open-letter-trump [https://perma.cc/6VSU-WWFB]; Cecilia Kang & Kate Conger, *Snap Says It Will No Longer Promote Trump's Account*, N.Y. Times (June 3, 2020), https://www.nytimes.com/2020/06/03/technology/snapchat-trump.html [https://perma.cc/R7R7-4V5A]; Cat Zakrzewski, *The Technology 202: Protests, Coronavirus and Election Present Disinformation for Challenge Social Media Companies*, Wash. Post. (June 1, 2020), https://www.washingtonpost.com/news/powerpost/paloma/the-technology-202/2020/06/01/the-technology-202-protests-coronavirus-and-election-present-disinformation-challenge-for-social-media-companies/5ed3f3a9602ff12947e801f0/ [https://perma.cc/8T36-AX64] (explaining that tech companies generally respond in unison to difficult content postings but do not in regards to the U.S. president's tweets).

[265] *See supra* Part I.

[266] *See* Molly K. Land & Rebecca J. Hamilton, *Beyond Takedown: Expanding the Toolkit for Responding to Online Hate*, *in* Propaganda and International Criminal Law: From Cognition to Criminality 143 (Predrag Dojcinovic ed. 2020).

To the degree that this insight is accurate, it has implications for the utility of the Oversight Board that Facebook introduced in 2019.[267] It may be that improved content regulation requires not simply better content moderation decisions—something on which the Oversight Board is set to work—but structural changes that cut to the very heart of the business model on which Facebook and other social media platforms rely.

3. *The Global Public Square's Power Imbalance*

The function-based approach highlights that, in addition to states and technology companies, a variety of other actors, such as individual users, nonstate groups, international organizations, and traditional media, can work on rule creation. Yet, it also draws attention to the fact that only states and technology companies have the possibility of exercising direct enforcement.[268] There is, therefore, a mismatch of actors as between the rule creation and enforcement points of the model, which in turn creates a significant power imbalance to the detriment of those who are unable to take direct enforcement action.

The greater the alignment of interests between the actor seeking the removal of content, and the state or technology company who can remove it, the less this power imbalance matters. So, for example, even though the average individual user does not have much power within this system, if she seeks the removal of ISIS content, she is likely to find success given the interest technology companies have in keeping such content offline (both for inherent reasons and because of pressure from powerful states on this issue). This interest leads the technology company to both proactively monitor for ISIS content, investing in human and algorithmic training to this end, and react quickly when such content is brought to their attention. Work the same issues through in relation to getting casteist content removed in India, or anti-Rohingya content removed in Myanmar, however, and we see how challenging it is for even an organized group of users to get the changes they seek.

The function-based model's disaggregation of rule creation from enforcement also brings to light the power dynamics that influence whether an effort at collateral censorship will be successful. One might compare, for example, the ability of the German government arguing a piece of anti-Semitic content should be removed and the Sri Lankan government arguing a piece of anti-Sinhalese content should be removed. The power of the German government to fine a technology company readily aligns that company's interests with the interest the German government has in seeing the

[267] Evelyn Douek, *Facebook's "Oversight Board:" Move Fast with Stable Infrastructure and Humility*, 21 N.C. J. L. & Tech. 1 (2019).

[268] Save the exception, noted above, of hackers who can launch a DDoS attack to bring down a website directly. *Supra* note 253.

content removed. Thus collateral censorship is possible.[269] The same cannot be said for the Sri Lankan government which, lacking the power to secure the technology company's attention, is left with the choice of having the content remain online, or shutting down access to the platform entirely. The latter option leads the Sri Lankan government to violate the free expression rights of the entire Sri Lankan population in order to stop the flow of just one particular piece of content.[270]

4. *Reckoning with the Normative Weakness of the Current System*

The rule creation point of the function-based model shows pluralism in action. Different actors, each asserting that their preferred rule should be applied to the content, jostle for position. As in any pluralist system, "there is no external position from which one could make a definitive statement as to who is authorized to make decisions in any given case. Rather, a statement of authority is itself inevitably open to contest."[271]

This reflects exactly the situation we see playing out. Technology companies update their Community Guidelines in response to user feedback, as we saw with Facebook's new rule prohibiting casteist content following feedback from users in India.[272] States and technology companies negotiate and re-negotiate what rules apply in any given situation, as observed in the interactions between the Turkish government and Twitter.

The resulting indeterminacy is frustrating for many, but is also exactly what we should expect from the current system unless and until consensus can be built up on the question of what rules should apply. This then highlights a normative question for all those involved in the global public square: Is this pluralist approach to rule creation what we are comfortable with? Some may argue that such fluidity, beyond being descriptively accurate, has value in a moment where there is no consensus across those participating in the global public square over whose rules should apply. Others will view this in a negative light, noting that such fluidity will invariably lead to the outcomes we currently see, where those with political and financial power are more likely to get the enforcement action they seek. Looking to the future, there may be a third way that involves building consensus around some baseline norms, while preserving fluidity above that baseline to

[269] As Citron notes, "Tech companies yielded to European regulators' demands because they knew their threats were not idle." Citron, *supra* note 65, at 1046.

[270] As commentators have noted, shutting access to a platform is an overbroad measure that, on balance, is likely to do more harm than good. *See, e.g.*, Louise Matsakis & Issie Lapowsky, *Don't Praise the Sri Lankan Government for Blocking Facebook*, WIRED (Apr. 23, 2019), https://www.wired.com/story/sri-lanka-bombings-social-media-shutdown [https://perma.cc/EXN6-7G7B] (highlighting the problems arising from the government's decision to block Facebook after the Easter bombings in 2019).

[271] Paul Schiff Berman, *Global Legal Pluralism*, 80 S. CAL. L. REV. 1155, 1166 (2007).

[272] *See supra* Part II.A.

account for local difference, in other words, the integration of a "margin of appreciation" into platform governance.[273]

The outsized power of technology companies in this space is something that policymakers and the public are belatedly coming to understand. There is growing discomfort with the status quo where complex speech decisions are left in the hands of private actors.[274] The case studies here suggest that just as the triadic literature developed blind spots by failing to seek out a diversity of examples, any successful resolution of these normative questions will need to take account of the array of legal, social, and cultural fora in which these platforms operate.

Conclusion

I am not underestimating the challenges facing those involved in governing the global public square. As some high-profile failures have demonstrated, technology companies do not yet have the capacity through either human resources or AI to respond to the scale of speech complaints they receive on an hourly basis.[275] Even if or when they overcome these problems of scale, there may still be instances where they make bad speech decisions[276] and certainly they will almost never be able to make decisions that satisfy all concerned.

Part III.B of this Article provided concrete suggestions for improving the governance of the global public square as currently configured but it may ultimately be that such management "tweaks" are insufficient. As the function-based model highlights, it may be impossible to address some of the most pressing problems of content moderation without changing the underlying business model on which social media platforms operate.

Moreover, even if the platforms that I focused on in this Article make significant improvements, this will not be enough. One looming threat is that as major platforms like Facebook, YouTube and Twitter improve their ability to remove or de-emphasize dangerous speech, those who promulgate such content will simply move to non-public platforms, a trend already well underway.[277] Another is that Chinese state-run alternatives to the Silicon

[273] *See* Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 Fordham Int'l L.J. 939 (2020).

[274] Daphne Keller, *Internet Platforms: Observations on Speech, Danger, and Money,* Hoover Institution (2018), https://www.hoover.org/sites/default/files/research/docs/keller_webreadypdf_final.pdf [https://perma.cc/GWU7-TVRJ] (describing "a set of fundamental problems with delegating complex decisions about free expression and the law to private companies.")

[275] *See generally* Gillespie, *supra* note 67.

[276] *See, e.g.*, Terje Solsvik & Yasmeen Abutaleb, *Facebook Reinstates Vietnam Photo after Outcry over Censorship,* Reuters (Sept. 9, 2016), https://www.reuters.com/article/us-facebook-norway-primeminister/facebook-reinstates-vietnam-photo-after-outcry-over-censorship-idUSKCN11F194 [https://perma.cc/VXS3-XEUC].

[277] Already the spread of misinformation on private messaging systems such as Telegram and WhatsApp, where end-to-end encryption makes any kind of moderation near impossible, is rampant. *See, e.g.*,

Valley-based platforms succeed not only in China, but in many other states in which the government and/or users themselves are dissatisfied with the standards that U.S. companies impose. Nonetheless, no one, including the major technology companies themselves, would suggest that there is not plenty of scope for improvement on the current approach.

The core of this Article has focused on expanding the descriptive account of regulatory activity in the global public square that actually exists, rather than the narrowly circumscribed version that appears throughout the triadic literature. The accuracy of the account presented here matters, since there can be no hope of addressing the problems that everyone seems to acknowledge the system is producing without first understanding the true character of the system itself.

Luca Belli, *WhatsApp Skewed Brazilian Election, Proving Social Media's Danger to Democracy,* CONVERSATION (Dec. 5, 2019), https://theconversation.com/whatsapp-skewed-brazilian-election-proving-social-medias-danger-to-democracy-106476.5 [https://perma.cc/RGN9-KTZG].