The Governors' Advisors: Experts and Expertise as Platform Governance

Amre Metwally[*]

introduction

The United Kingdom's draft Online Safety Bill (OSB),[1] published in May of last year, generated a fresh round of criticism and scrutiny. The OSB seeks to delineate standards for companies in their moderation of lawful but ultimately harmful content, with the UK's regulatory body Ofcom empowered to levy fines, pressure platforms to improve their moderation efforts, and block sites that fail to comply with the regulation.[2] Though many voices in the online child safety arena applauded the effort,[3] other actors across civil society issued grave warnings. Big

[*] Juris Doctor Candidate, Harvard Law School. The author was previously a Policy and Enforcement Manager covering political extremism, counterterrorism, and graphic violence for YouTube. The author would like to thank Tomás Guarna for his incisive feedback on an earlier draft of the piece and Mehtab Khan, Brianna Yang, and the staff of the Yale Information Society Project for all their assistance in the editing process.

[1] The Bill was previously known as the Online Harms Bill. The journey to this final stage began in 2019 with the UK's white paper proposing a duty of care for social media platforms. *See generally Online Harms White Paper*, SEC'Y OF STATE FOR DIG., CULTURE, MEDIA & SPORT & THE SEC'Y OF STATE FOR THE HOME DEP'T (2020), https://www.gov.uk/government/consultations/online-harms-white-paper/online-harms-white-paper#executive-summary.

[2] *World-First Online Safety Laws Introduced in Parliament*, DEP'T FOR DIGITAL, CULTURE, MEDIA & SPORT (Mar. 17, 2022), https://www.gov.uk/government/news/world-first-online-safety-laws-introduced-in-parliament.

[3] Edina Harbinja, *U.K.'s Online Safety Bill: Not That Safe, After All?*, LAWFARE (July 8, 2021, 1:36 PM), https://www.lawfareblog.com/uks-online-safety-bill-not-safe-after-all.

Brother Watch argued that the Bill "introduces state-backed censorship and monitoring on a scale never before seen in a liberal democracy."[4] The fear of censorship is attributable to the government's crackdown on "vague categories of lawful speech."[5] Others have also warned that the duty of care to be established in the OSB fails to embrace key principles, such as appropriate notice and appeals for users, and could unduly burden smaller platforms as well.[6]

The UK model reflects only one governance approach to handling social media platforms. Other governance models have also surfaced throughout the last decade to address platform regulation more broadly[7] or to home in on particular policy areas of public concern more specifically.[8] At the same time, companies themselves have rushed to explain how they have worked to self-regulate, improve, enhance, develop, and grow their abuse-fighting teams to promote a healthy internet.

No matter the governance approach, whether one examines the OSB, the EU's Digital Services Act, or other regulatory proposals, an elusive

[4] Big Brother Watch Team, *Big Brother Watch Response to the Government's Online Safety Bill*, BIG BROTHER WATCH (May 12, 2021), https://bigbrotherwatch.org.uk/2021/05/big-brother-watch-response-to-the-governments-online-safety-bill.

[5] *Id.*

[6] Christoph Schmon, *UK's Draft Online Safety Bill Raises Serious Concerns Around Freedom of Expression*, ELEC. FRONTIER FOUND. (July 14, 2021), https://www.eff.org/deeplinks/2021/07/uks-draft-online-safety-bill-raises-serious-concerns-around-freedom-expression.

[7] *See, e.g.*, *Commission Proposal for a Regulation of the European Parliament and of the Council on a Single Market for Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, COM (2020) 825 final (Dec. 15, 2020), https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52020PC0825&from=en (establishing a regulatory regime for online intermediaries in general, from social media platforms to app stores, and online marketplaces); *see also* Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [NetzDG] [Network Enforcement Act], Sept. 1, 2017, BUNDESGESETZBLATT (utilizing existing domestic law as the basis for companies to respond to and remove content that violate German law), Teil I [BGBL I] (Ger.); Türk Medeni Kanunu, Kanun No: 7253 R.G.: 31.07.2020 Sayı 31202, Kabul Tarihi: 29.07.2020 (Tur.) (requiring social media platforms to appoint a local representative to respond to court orders from Turkey for content removals, with non-compliance resulting in an effective ban of the site by severely limiting internet traffic to the site); Information Technology (Intermediary Guidelines and Digital Media Ethics Code) Rules, 2021 (India) (implementing a governance structure similar to Turkey's approach for platforms to respond to unlawful content within 24 hours and also publish a monthly compliance report).

[8] *See, e.g.*, Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [NetzDG] [Network Enforcement Act], Sept. 1, 2017, BUNDESGESETZBLATT, Teil I [BGBL I] (Ger.) (imposing stringent requirements for companies to comply with content removals that run afoul of German law, with a particular focus on hate speech).

term appears repeatedly in the texts: "expertise." The UK, for example, envisions a platform's duty of care for users to be tied to a code of practice that details what steps a company should take to comply. In an interim code of practice for terrorist content, the UK instructs companies to "commit to collaborative working with industry and with governments, academia and civil society" and to "engage with relevant industry bodies, which enable the sharing of knowledge and expertise."[9]

At the same time, this term is invoked by the decisionmakers themselves, these "New Governors"[10] that write and enforce the rules of speech on social media platforms. YouTube's Community Guidelines page on misinformation notes that the company's policies "are developed in partnership with a wide range of external experts."[11] As part of an update in 2019 on the video-sharing network's effort to respond to hateful speech, a company blog post boasts that several updates were made after consulting "dozens of experts in subjects like violent extremism, supremacism, civil rights, and free speech."[12]

The natural question, then, is: who are these experts? And how is their expertise deployed to cajole platforms into specific ways of thinking about and censoring online speech? While Kate Klonick and others have discussed the ways in which "outside influence" aids in the platforms' efforts to iterate on their policies,[13] expert factions engage in a more assertive process of worldmaking by exerting control without proper disclosures from the social media companies that work with them. This paper argues that expert communities themselves have emerged as a mode of platform governance—employing a range of techniques to advise and demand social media networks to embrace the norms, policy ideas, and solutions they hold most dearly. The diversity and pluralism, in other words,

[9] *Interim Code of Practice on Terrorist Content and Activity Online*, DEPT. FOR DIG., CULTURE, MEDIA & SPORT, (2020), https://www.gov.uk/government/publications/online-harms-interim-codes-of-practice/interim-code-of-practice-on-terrorist-content-and-activity-online-accessible-version#section-2-collaboration [hereinafter *Interim Code*].

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

[11] *How Does YouTube Address Misinformation?*, YOUTUBE, https://www.youtube.com/howyoutubeworks/our-commitments/fighting-misinformation/ (last visited Jan. 9, 2022).

[12] *Our Ongoing Work to Tackle Hate*, YOUTUBE (June 5, 2019), https://blog.youtube/news-and-events/our-ongoing-work-to-tackle-hate.

[13] *See* Klonick, *supra* note 10, at 1648-49 (outlining four ways that content moderation policies are subject to outside influence).

necessary to make social media platforms reflective of the world can also become tools to constrain, harden, or impose certain world views.

The paper discusses the types of experts engaged in platform governance before analyzing the tactics and methods at their disposal. The central contention in this paper is that the platforms themselves are the "foreground deciders"[14] and that these experts, or "people with projects,"[15] operate in the background, instead "advis[ing] and interpret[ing] by inhabiting modes of knowledge and communication through which they can pursue projects with some plausible deniability of agency."[16] The ideological agendas and facts that serve as a basis for a preferred vision of platform governance are socially constructed.[17]

I. Who are the Experts?

The array of experts in and around social media companies can be dizzying. While a proper cartography of all the actors is needed, this paper does not promise this. Instead, in this Part, I put forth broad categories that situate the experts based on their positionality, whether inside or outside the social media companies, as members of civil society groups, or in the emerging Trust & Safety industry more broadly. The expertise held by different expert types can work in tandem across boundaries, creating or maintaining certain worldviews. These expert ecosystems can also bring forward expertise and knowledge that clashes with the agenda or ideology of another ecosystem. This mapping does not suggest that there is uniformity of thought across these actors but rather that each faction possesses certain worldviews and goals as they work to lobby, influence, and advise a platform. Part II discusses the governance techniques utilized to advance visions and ideas that a particular expert group may possess.

A. Outside Experts

Outside experts are the third-party consultants, research institutes, and companies that offer formal services—often for pay—on particular topics for social media companies. When it comes to combatting "terrorism,"

[14] DAVID KENNEDY, A WORLD OF STRUGGLE: HOW POWER, LAW, AND EXPERTISE SHAPE GLOBAL POLITICAL ECONOMY 111 (2018)

[15] *Id.*

[16] *Id.*

[17] *See id.* at 112.

most, if not all, social media platforms have at least some mention of a prohibition of terrorist content. What sources, however, do companies rely on to confer this terrorist designation on their content and users? Most companies remain tight-lipped on the contents of the sources that guide their decision-making, though stellar reporting from *The Intercept* helps shed some light. Their reporters published Facebook's Dangerous Individuals and Organizations list in its entirety,[18] which included reference to the source(s) used to decide an entity or individual addition to the platform's list. While some sources, such as the United States government's Specially Designated Global Terrorists list, may not elicit shock, Facebook's list also included the "Terrorism Research & Analysis Consortium, a private subscription-based database . . . and SITE, a private terror-tracking operation with a long, controversial history."[19]

In an old blogpost written by Facebook executives, the company itself discloses some of its "partners"—such as Flashpoint, the Middle East Media Research Institute (MEMRI), and SITE—who provide "expertise in global terrorism or cyber intelligence."[20] It is precisely this ecosystem of academics, research centers, think tanks, and private "risk intelligence" companies that form a web of "outside experts."[21] It is these third-party actors that companies often flaunt when they vaguely mention "experts" and "knowledge" that help shape their policy changes or enforcement approaches.

Relying on the use of outside experts may not seem concerning; however, the problems of relying on these third-party players stem from a genuine lack of transparency as to the breadth and depth of control or influence an outside expert has over a platform. Does a private risk monitoring service, for example, offer its services to find offensive content

[18] Sam Biddle, *Revealed: Facebook's Secret Blacklist of "Dangerous Individuals and Organizations,"* THE INTERCEPT (Oct. 12, 2021, 1:16 PM), https://theintercept.com/2021/10/12/facebook-secret-blacklist-dangerous.

[19] *Id.*

[20] Monika Bickert & Brian Fishman, *Hard Questions: Are We Winning the War on Terrorism Online?*, FACEBOOK NEWSROOM (Nov. 28, 2017), https://about.fb.com/news/2017/11/hard-questions-are-we-winning-the-war-on-terrorism-online.

[21] For a discussion specifically about "outside experts" in the counterterrorism content moderation industry, *see* Amre Metwally, *"Outside Experts": Expertise and the Counterterrorism Industry in Social Media Content Moderation*, 12 J. NAT'L SEC. L. & POL'Y 471 (forthcoming 2022). *See also* Klonick, *supra* note 10, at 1655 ("For a number of years, platforms have worked with outside groups to discuss how best to construct content-moderation policies . . . third-party groups had and continue to have an impact on the policies and practices of major social media platforms.").

based on its own criteria or a platform's (and moreover, what are those criteria?)? Will an academic aid in formulating part of a company's policy, or shape the overall posture for an issue area?

Most importantly, this lack of transparency robs us all of the ability to know *which* actors—these people with projects—are listened to and which ones are ignored, or which worldviews a company embraces and why. In framing this work with outside experts as ostensibly neutral acts of consultations or advice, companies obfuscate the fact that each entity has its own goal and agenda, further complicated by the fact that many of these entities offer their "expertise" for a fee. While the platform ultimately retains the discretion in the design of its content policies, a company also has the discretion to choose *which* outside experts with which to engage and which to ignore. For example, Facebook has been pressured by the International Holocaust Remembrance Alliance to include the term "Zionist" as a racial slur per its hate speech policies,[22] despite the fact that many other outside experts have expressed such a posture as "highly problematic."[23] The backgrounds, training, personal and political ideologies of the actor matter significantly in determining the "modes of knowledge"[24] that are advertised and sold to technology firms.

B. Inside Experts

Experts inside companies are the policy and engineering teams responsible for ensuring user safety. They are the employees responsible for developing and modifying content policies and designing the detection algorithms and enforcement apparatus to review violative material. As individuals, these employees have their own biases and their own discretion that can sway policy choices, particularly as they advise the company's leadership who ultimately bear the public's wrath for a given decision. Whether we consider companies as anthropomorphized beings or scrutinize a firm's leadership, the decisions and choices from internal teams are ultimately *not* the face of a platform or its leadership staff. Tellingly, *The New York Times*, in a piece on YouTube's CEO Susan

[22] Jillian York & David Greene, *Facebook's Latest Proposed Policy Change Exemplifies the Trouble with Moderating Speech at Scale*, ELEC. FRONTIER FOUND. (Feb. 4, 2021), https://www.eff.org/deeplinks/2021/02/facebooks-latest-proposed-policy-change-exemplifies-trouble-moderating-speech-0.

[23] Amos Goldberg, *Dear Facebook: Please Don't Adopt the IHRA Definition of Antisemitism*, FORWARD (Sept. 13, 2020), https://forward.com/opinion/454124/dear-facebook-please-dont-adopt-the-ihra-definition-of-antisemitism.

[24] *See* KENNEDY, *supra* note 14, at 111.

Wojcicki, implicitly reflects this reality as well, noting that "Ms. Wojcicki said she know that **her** policy changes could 'upset some people.'"[25] Ultimately Ms. Wojcicki and her senior executives are the foreground deciders on policy, while the technocratic internal teams work diligently in the background to consider the array of policy, design, and engineering changes. After all, even from my time at YouTube, unpopular decisions made by members of my team never generated calls for me, my peers, or my colleagues to be held accountable. Even in recent disclosures such as the Facebook Files,[26] internal experts are the first to maintain deniability, pointing blame at the company leadership directly.[27]

This naïve reading, however, maintains a narrative of neutral, objective technocracy. Employees themselves have their own discretion, opinions, backgrounds, and worldviews. I, for example, saw certain topics very differently from my teammates—who wins and who loses? Who has the "sway" to propose a path forward before this proposal snakes its way up the company ladder for approval? Each choice to, say, narrowly define "terrorism" or widen its scope or develop a classifier that targets hate speech in one language but not another is the by-product not only of the foreground deciders' blessing but also one of internal wrangling. Some background voices are louder than others. Even in an effort to identify and hire candidates for policy teams, companies have become increasingly vocal about who they want operating in these policy spheres to formulate the approaches that are ultimately presented to the foreground actors. By framing job descriptions in language such as "qualifications" and years of experience in government, civil society, or at other policy teams inside the technology industry, companies have increasingly begun equating specialized knowledge in a particular subject area as an opportunity to promote certain narratives over others.[28]

[25] Daisuke Wakabayashi, *The Most Measured Person in Tech is Running the Most Chaotic Place on the Internet*, N.Y. TIMES (Apr. 17, 2019) (emphasis added), https://www.nytimes.com/2019/04/17/business/youtube-ceo-susan-wojcicki.html.

[26] *See generally The Facebook Files*, WALL ST. J. (2021), https://www.wsj.com/articles/the-facebook-files-11631713039 (disclosing internal documents that show, among other things, employee presentations to company executives of policy problems plaguing the site).

[27] *See, e.g.*, Mark Bergen, *YouTube Executives Ignored Warnings, Letting Toxic Videos Run Rampant*, BLOOMBERG NEWS (Apr. 2, 2019, 11:29 AM), https://www.bloomberg.com/news/features/2019-04-02/youtube-executives-ignored-warnings-letting-toxic-videos-run-rampant (discussing how employee recommendations to deal with toxic content was allegedly ignored by YouTube's senior leadership).

[28] I examine this point within the context of terrorist content moderation and former terrorism experts hired as employees in the technology sector. *See* Metwally, *supra* note 21, at 482-485

C. *Trust & Safety Institutionalization: TSPA, Integrity Institute*

Over the last two decades, content policy and enforcement work has bloomed into an established industry. Also referred to as "Trust & Safety" or "Integrity" functions, this sector has become non-negotiable in many companies that design platforms that facilitate user-generated content, connect people on or offline, or allow for the exchange of goods or services. As this industry has become more established, we are beginning to see individuals who have moved through the ranks of this nascent field and become increasingly vocal, particularly after leaving the sector. In recent years, former Trust & Safety professionals have become active in organizing their own initiatives, groups, and organizations that bring together knowledge from professionals in this field—notably the Trust & Safety Professional Association[29] and the Integrity Institute.[30]

Although these groups are often comprised of former Trust & Safety employees, they deserve their own categorization due to the fact they operate without profit and do not work directly with, or on behalf of, the technology companies. These organizations are the latest evolution in a movement to professionalize and institutionalize the knowledge, skills, and work that encompass Trust & Safety. While a company may hire individuals who possess sets of knowledge—hate speech, misinformation, intelligence, for example—this shift centers the work and expertise of Trust & Safety as a body of knowledge itself.

Both the TSPA and Integrity Institute produce material for the public. The TSPA, however, also focuses on the Trust & Safety community. For example, the organization created and published a Trust & Safety curriculum that identifies the "core concepts, terms, and standard practices that make up the body of knowledge we call 'trust and safety.'"[31] The Trust & Safety professional, in other words, carries expertise and knowledge no matter the company in which they work, and this curriculum is a reflection of a need, or desire, to standardize the processes, work, and output a professional in this space is expected to produce. The TSPA also

(considering Facebook's decision to hire Brian Fishman to run its counterterrorism team as well as YouTube's effort to build an "Intelligence Desk" to monitor for new threats, with these job postings often coveting experience in intelligence-related work).

[29] *See generally* TRUST & SAFETY PRO. ASS'N, https://www.tspa.org/ (last visited Jan. 12, 2022).

[30] *See generally* INTEGRITY INST., https://integrityinstitute.org/ (last visited Jan. 12, 2022). The author is also a member of the Integrity Institute.

[31] *Trust & Safety Curriculum*, TRUST & SAFETY PRO. ASS'N, https://www.tspa.org/curriculum/ts-curriculum/ (last visited Jan. 12, 2022).

aggregates material for Trust & Safety professionals through its resource library, which is a space "to collect articles, papers, lectures, and podcasts that trust and safety professionals may find useful in developing policies, supporting moderators, building systems to detect violations, and generally deepening their practice."[32] Though the TSPA does not engage in more aggressive professional governance of Trust & Safety professionals, these efforts reflect a desire to situate the work of Trust & Safety outside of any one company and perhaps a form of softer governance approach to the methods of carrying out this type of work.

The Integrity Institute, on the other hand, views itself as a conduit for its members—many of whom are former employees at social media companies—to join the public debates about integrity work at social media companies. Its goal is for its members' "experience [to] be put to use for the social good and have impact in the public conversation."[33] Through advising companies facing integrity-related problems,[34] meeting with policymakers or journalists,[35] or producing reports and recommendations on a variety of integrity-related questions (for example metrics and transparency, or ranking algorithms),[36] the Integrity Institute also strives to centralize integrity knowledge as a discipline of its own. This expertise, and the accompanying integrity-as-knowledge work, does not "just generate knowledge; it determines legal and policy decisions"[37] or, framed another way, joins the "public conversation."[38]

D. Civil Society

Civil society groups, including non-governmental organizations and advocacy groups, all over the world have emerged as increasingly powerful expert bodies in platform governance. With countless examples of

[32] *Resource Library*, TRUST & SAFETY PRO. ASS'N, https://www.tspa.org/explore/resource-library/ (last visited Jan. 11, 2022).

[33] *About Us*, TRUST & SAFETY PRO. ASS'N, https://integrityinstitute.org/home#about (last visited Jan. 10, 2022).

[34] *See id.*

[35] *See id.*

[36] *Integrity Institute Resources*, INTEGRITY INST., https://integrityinstitute.org/resources (last visited Jan. 13, 2022).

[37] Sanne Taekma, *Expert Accountability and the Rule of Law: Intertwinement of Normative and Functional Standards?*, *in* TECHNOCRACY AND THE LAW: ACCOUNTABILITY, GOVERNANCE AND EXPERTISE 45 (Alessandra Arcuri & Florin Coman-Kund eds., 2021).

[38] *About Us*, INTEGRITY INST., https://integrityinstitute.org/home#about (last visited Jan. 10, 2022).

companies, their policies, and their algorithms failing to capture linguistic, cultural, and regional complexity in their content moderation enterprise,[39] civil society groups have stepped in to fulfill several roles. First, these actors have started to play a larger role in the policy creation process, as evidenced by civil society advisory boards touted by TikTok, Twitter, and others.[40] The depth and efficacy of their involvement, however, remains highly dependent on the design of the companies' stakeholder engagement process. For example, as Brenda Dvoskin notes, the input itself is "solicited to specific stakeholders"[41] and that "means that access is heavily controlled by the company."[42]

While some actors more tightly control the content moderation inputs from civil society actors, other bodies, notably Facebook's Oversight Board, has designed an input process that allows all concerned voices to weigh in on a particular topic.[43] The Board itself can also issue policy statements that "will be taken into consideration by Facebook to guide its future policy development."[44] Though not all of the Board's members are from civil society, a significant number are, creating a promising avenue to ensure at least some voices from this sphere are represented in a powerful forum. Finally, civil society actors have also played critical roles in norm-setting, which will be explored in more detail in Part II.

E. Companies in the Content Moderation Enterprise

[39] *See, e.g.*, Michael Levenson, *Instagram Blocked Posts about the Aqsa Mosque in a Terrorism Screening Error*, N.Y. TIMES (May 13, 2021), https://www.nytimes.com/2021/05/13/world/middleeast/instagram-aqsa-mosque.html; Richard Ashby & Molly Land, *Hate Speech on Social Media: Content Moderation in Context*, 52 CONN. L. REV. 1029, 1068 (2021) (discussing the use of hate speech slur lists maintained by social media companies and the ways this approach has failed in Myanmar and the ongoing genocide against its Rohingya community); Spandana Singh & Eliza Campbell, *Content Moderation Trends in the MENA Region: Censorship, Discrimination by Design, and Linguistic Challenges*, NEW AM. (Aug. 25, 2021), https://www.newamerica.org/oti/blog/content-moderation-trends-in-the-mena-region-censorship-discrimination-by-design-and-linguistic-challenges ("Similarly, a member of civil society also noted during our interviews that a Saudi-Arabic-speaking user had their post on Twitter referring to a goal in a soccer match removed, likely because the colloquial word for goal in his dialect roughly translates to 'missile.'").

[40] Brenda Dvoskin, *Representation Without Elections: Civil Society Participation as a Remedy for the Democratic Deficits of Online Speech Governance*, 14 VILL. L. REV. (forthcoming 2022), https://ssrn.com/abstract=3986181.

[41] *Id.*

[42] *Id.*

[43] *Id.*

[44] *Oversight Board Charter*, art. 3, § 4, META (Sept. 2019), https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf.

There are dozens, if not hundreds, of platforms offering user-generated content or other forms of social connectivity. While one may consider any single company as a foreground decision maker, it is important to consider the ecosystem of "peers" that operate as background experts and influencers. While companies have said before that each maintain their own policies, the actions and choices of other platforms sway a foreground actor. For example, in its reporting on Facebook's Dangerous Individuals and Organizations list, *The Intercept* noted that it "appears Facebook has worked with its tech giant competitors to compile the DIO list; one entry carried a note that this entity had been "escalated by" a high-ranking staffer at Google who previously worked in the executive branch on issues related to terrorism. (Facebook said it does not collaborate with other tech companies on its lists.)"[45] Companies have also turned to industry-wide efforts to explore specific policy issues such as online terrorist content, and these programs complicate our understandings of bespoke policy and enforcement approaches for each platform.[46]

Additionally, companies have routinely taken decisions shaped partially in response to the decisions of their peer institutions. For example, in only a matter of hours from each other, Apple, Spotify, YouTube, and Facebook all moved to remove Alex Jones from their platforms for violating their Terms of Service.[47] Jones, who often propagates deeply offensive conspiracy theories and incendiary racist rhetoric, had been permitted to use these services for years before Apple's first step—after increasing public criticism[48]—triggered his swift deplatforming. While conservative voices may point to this event, or the situation regarding Donald Trump's Twitter account, to argue there is evidence of "collusion," the story is more complicated.[49] As Yochai Benkler notes in his

[45] Biddle, *supra* note 18.

[46] *See infra* Part II.

[47] Aja Romano, *Apple Banned Alex Jones's Infowars. Then the Dominoes Started to Fall*, VOX (Aug. 6, 2018, 6:18 PM), https://www.vox.com/policy-and-politics/2018/8/6/17655516/infowars-ban-apple-youtube-facebook-spotify.

[48] *Id.*

[49] In perhaps a form of collaboration that might edge closer towards "collusion," platforms have begun creating industry initiatives such as the Global Internet Forum to Combat Terrorism (GIFCT) and the Digital Trust and Safety Partnership (DTSP), among other programs. *See generally* GLOB. INTERNET F. TO COMBAT TERRORISM, https://gifct.org/ (last visited Jan. 13, 2022); DIGIT. TRUST & SAFETY P'SHIP, https://dtspartnership.org/ (last visited Jan. 13, 2022). This will be scrutinized in Part II as a governance technique that, I argue, intends to further a goal to standardize policy and enforcement across platforms.

examination of the attack on Wikileaks from both public and private actors, their movements were less "a single coordinated response" but rather "a series of acts . . . that feed into each other."[50] Companies routinely assert the need to protect users' freedom of expression but that line is "an increasingly handy excuse"[51] to circumvent responsibility. Once the first platform made a more aggressive maneuver, then "the more controversial action would have been to allow Jones and Infowars to remain. And so, sites that just a week ago were tentatively committed to protecting Jones's 'free speech' couldn't about-face fast enough."[52] It is precisely this type of cross-company influencing where we see the "deniability" of background experts at play: companies that choose *not* to operate in lockstep with its peers would then have the unenviable task of explaining themselves. It is much easier to avoid explanations of substance with the sound of dozens of doors closing shut; the voice that remains echoes much more loudly.

II. How and why do the Advisors Govern?

Though each category of experts may have different agendas and ideologies, they all are engaged in an effort to influence, shape, and thereby control, how the decision makers view and respond to content policy and enforcement challenges on the platforms. This Part examines the governance tools these experts often employ to assert their expertise. I argue that the methods are all in the furtherance of four main goals: first, re-imagining the "rules of the road" by engaging in norm-making and narrative-setting; second, shaping the content policies themselves; third, shaping the enforcement of discrete pieces of content; and fourth, standardizing the Trust & Safety enterprise across companies.[53] Each of these tactics are advanced through specific governance approaches that these experts use—though not every expert category may use each of these strategies.

A. Norm-Making and Narrative-Setting

[50] Yochai Benkler, *A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate*, 46 HARV. C.R.-C.L. L. REV. 311, 330-31 (2011).

[51] *Id.*

[52] *Id.*

[53] This last goal, as will be discussed in more detail in this section, is only relevant for the last category of experts listed in Part I: the companies.

One of the most effective ways experts can leverage power is through challenging the status quo—whether by offering an alternative narrative or presenting a new way of conducting content moderation. Outside experts, for example, routinely take to public outlets to voice frustration or even praise the platforms' work. Rita Katz, founder of SITE Intelligence Group, penned an opinion piece sharing her opinion, and findings of a study her organization carried out, that YouTube had become an example of how to respond to terrorist content on a social media platform.[54] Tellingly, after she presented her own evaluation of YouTube's performance, she also used the piece to put forward a norm that "[s]tifling terrorist propagandists and recruiters demands a far more collaborative, coordinated approach between governments, tech companies, and third-party entities."[55]

Other outside experts turn to the public eye to criticize the companies—but even criticism can be a powerful force to make platforms respond in ways that outside experts see fit. For example, during a hailstorm of criticism and fury that YouTube had allowed ISIS content to proliferate on its platform, the Counter Extremism Project (CEP) began raising numerous criticisms of YouTube's failure to censor sermons and content from Anwar al-Awlaki, an imam and key organizer for al-Qaeda.[56] Three months after CEP's vocal condemnations and a report of its own, YouTube finally began removing Awlaki content from its platform, ultimately generating praise from CEP as a "watershed moment."[57]

[54] *See* Rita Katz, *To Curb Terrorist Propaganda Online, Look to YouTube. No, Really.*, WIRED (Oct. 20, 2018, 8:00 AM), https://www.wired.com/story/to-curb-terrorist-propaganda-online-look-to-youtube-no-really; *see also* Metwally, *supra* note 21, at 499-500 (discussing Rita Katz's op-ed piece in *WIRED* in which she highlights her organization's analysis of YouTube's terrorism moderation efforts as proof that the company has become a leader in content moderation in this space).

[55] *Id.*

[56] *See, e.g.*, Scott Shane, *Internet Firms Urged to Limit Work of Anwar al-Awlaki*, N.Y. TIMES (Dec. 18, 2015), https://www.nytimes.com/2015/12/19/us/politics/internet-firms-urged-to-limit-work-of-anwar-al-awlaki.html?_r=0. Anwar al-Awlaki was a key recruiter for Al-Qaeda and was also the first American citizen to be killed by an American drone strike as part of his involvement with the terrorist organization. *See also* Metwally, *supra* note 21, at 500-501 (highlighting the Counter Extremism Project's public commentary criticizing the presence of Awlaki content on social media platforms and YouTube's subsequent decision to remove all of his work from the platform).

[57] Scott Shane, *In "Watershed Moment," YouTube Blocks Extremist Cleric's Message*, N.Y. TIMES (Nov. 12, 2017), https://www.nytimes.com/2017/11/12/us/politics/youtube-terrorism-anwar-al-awlaki.html.

Trust & Safety organizations have also seen the potential of capitalizing on their former employees' expertise to try and establish new norms for companies that operate in this space. The Integrity Institute, for example, has already published two reports outlining what it and its members argue should be the requirements for transparency and metrics as well as algorithmic ranking practices.[58] In presenting what the Institute argues is the "consensus view of Integrity Professionals,"[59] the Institute's unique position as a voice of Trust & Safety experts lends weight and expertise to the public discussion.

Civil society organizations have long been active participants in this space. Using a variety of advocacy campaigns and their own thought leadership, these expert actors have attempted to establish sets of norms and best practices that companies should adhere to for transparency and accountability,[60] and on intermediary liability.[61] These actors also wield advocacy campaigns to attempt and influence the design of content policies, a topic that is explored in more detail in the following sub-section.

B. *Shaping the Content Policies*

Experts also want to influence the design of the content policies that guide what, how, and whether content is censored online. Inside experts, hired by companies for particular bodies of substantive knowledge—such as child safety, hate speech, elections, and so on—are of course the most active in this space. After all, company employees themselves have their own opinions, moral compasses, and agendas that undoubtedly influence *how* they view particular problems and questions that surface for a social media platform. However, inside and outside experts appear to collaborate on the design of policies, as evidenced through company blog posts,

[58] *See Metrics & Transparency: Data and Datasets to Track Harms, Design, and Process on Social Media Platforms*, INTEGRITY INST. (Sept. 22, 2021), https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834d31bcf2c5ac4c07494/1635267795944/Metrics+and+Transparency+-+Summary+%28EXTERNAL%29.pdf [hereinafter *Metrics & Transparency*]; *Ranking and Design Transparency: Data, Datasets, and Reports to Track Responsible Algorithmic and Platform Design*, INTEGRITY INST. (Sept. 28, 2021), https://static1.squarespace.com/static/614cbb3258c5c87026497577/t/617834ea6ee73c074427e415/1635267819444/Ranking+and+Design+Transparency+%28EXTERNAL%29.pdf.

[59] *Metrics & Transparency*, *supra* note 58.

[60] *See generally* SANTA CLARA PRINCIPLES, https://santaclaraprinciples.org/ (last visited Jan. 7, 2022).

[61] *See generally* MANILA PRINCIPLES ON INTERMEDIARY LIABILITY, https://manilaprinciples.org (last visited Jan. 6, 2022).

announcements, and interviews.[62] It is difficult to assess the extent of collaboration or level of influence one set of experts has over the other, however, since companies traditionally do not disclose details, choosing instead to refer to "external engagement" or "consultation" in the policy development process.

Facebook, for example, publishes minutes from its Product Policy team meetings which can help illustrate, to some degree, the types of involvement that may take place.[63] In meeting minutes from 13 July 2021, Facebook employees laid out recommendations on how the company's policy for attacks on public figures should provide additional protections for public figures against speech that could "sexualize, degrade them [public figures], or attack their appearance" without "censoring political or benign commentary."[64] The company notes that 79 external engagements took place but does not provide the names of the people or entities with which Facebook consulted, only listing broad categories such as "free speech advocates," "human rights experts" and "public figures."[65] Nor does the company list the countries these external experts are based in—a map is embedded in the report, but the manner in which locations are marked allow a reader to identify some nations more easily than others.[66] The presentation concludes with laying out sets of options for each category of attacks on public figures and, tellingly, the range of views from the experts mapped by least to maximum protection for public figures.[67]

While Facebook's level of transparency is a step in the right direction (and markedly more forthcoming compared to other large platforms), these reports fail to adequately provide the level of feedback outside engagement provided on particular policy options—and whether Facebook even provided these options or instead interacted with outside experts using other qualitative or quantitative methods. Regardless of the nature, and type, of engagement, these details are not an extravagance. For example, if the company approached an external expert without providing

[62] *See, e.g.*, *How Does YouTube Address Misinformation?*, *supra* note 12.

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

[64] *Recommendation: Attacks on Public Figures*, FACEBOOK 3 (Oct. 2021), https://about.fb.com/wp-content/uploads/2021/10/Facebook_PolicyForum_-Attacks-on-Public-Figures.pdf.

[65] *Id.* at 5.

[66] *Id.*

[67] *Id.* at 10.

details on its proposed policy approaches or asked questions in ways that framed problems in particular ways, these design choices drastically alter the efficacy and independence of these outside experts. Or, alternatively, if the conversations had taken place in a much more collaborative tone, more akin to "What would you do if you could write this policy?" then such engagement becomes less of a consultation and more of an invitation to steer Facebook employees towards certain desired outcomes in the eyes of an outside expert. Also, without knowing the identities of the actors who took part in these external engagements, we are unable to assess which viewpoints and voices were given credence and which ones, by the act of omission, were silenced.

These similar tensions are also present in the advisory roles civil society groups often hold for technology companies. While these sites of interaction provide *some* opportunity for *some* civil society actors to be in direct conversation with policy teams at companies, these boards, such as Twitter's Trust and Safety Council,[68] are ultimately designed and controlled by companies. Staff retain the ability to create such a board in the first place, invite the organizations they see most fit to participate, and determine the scope of the interactions.[69] Civil society experts may have more influence in shaping the policy may be in targeted advocacy and coalition-building work,[70] waging insurgent campaigns to shape policy through seizing media attention and public discourse and opinion.[71] Remaining confined to roles in advisory councils and boards established piecemeal by companies that choose to engage in this type of engagement

[68] *Trust and Safety Council*, TWITTER, https://about.twitter.com/en/our-priorities/healthy-conversations/trust-and-safety-council (last visited Apr. 3, 2022).

[69] *See* Dvoskin, *supra* note 40. Even in instances where a council or board purports to have complete independence, such as Meta's Oversight Board. Meta's staff still retains the discretion to implement policy recommendations as they see fit.

[70] For example, 7amleh, Social Media Exchange, the Council on American Islamic Relations, Kairos, and Eyewitness Palestine, among other partners, organized a campaign titled "Facebook, we need to talk." *See generally* FACEBOOK, WE NEED TO TALK, https://facebookweneedtotalk.org/ (last visited Jan. 4, 2022). The campaign, which included petitions, text for users to post on their own social media accounts, and events to challenge Facebook's decision to equate the term "Zionist" with a slur for the purposes of its hate speech policy. Many outside experts and civil society groups will also use report writing as a key method to lay out arguments for the design, expansion, or contraction of a policy. Risk intelligence companies, for example, often provide intelligence reports that highlight new types of problems, organizations, or threats for a particular sphere of policy. NGOs will publish reports that criticize a policy and call for reforms for a policy to be more equitable.

[71] *See* Klonick, *supra* note 10, at 1652 ("The media do not have a major role in changing platform policy per se, but when media coverage is coupled with . . . the collective action of users . . . platforms have historically been responsive.").

may be too restrictive or exclusionary for civil society groups seeking change.

C. Shaping the Enforcement

Policy and enforcement are intimately intertwined. After all, a platform's policy is only as good as its enforcement of ostensibly violative material—enforcement is what inevitably alters, modifies, and evolves the set of policies for a company. As such, it is important to remember that while an important outcome of expert-involved platform governance is to shape the policies, contesting content can be a critical avenue of governance. Each discrete tweet, post, or image is not only a site of contestation but also an iterative process of refining a platform's policies.

One interesting technique to advance this goal is in the flagging, or reporting, infrastructure embedded in many platforms. Traditionally, flagging has been a recourse available to a platform's users, inviting them into the moderation enterprise by allowing users to voice their concern for potentially objectionable material and notify companies that review may be necessary.[72] Flagging systems, as Klonick writes, are dependent on users, in part to "legitimize the system when platforms are questioned for censoring or banning content."[73] With time, other actors have been invited to participate in the moderation project, to legitimize the platforms' system, through flagging content as well. Outside experts—in particular, third-party "risk intelligence" companies that promise to help platforms fight terrorism, child abuse, hate speech, violence, and other ills—have embraced flagging as a key tool for their work.

Take, for example, ActiveFence, an Israeli firm that promises to prevent "evil at scale."[74] On its website, the company tells platforms they can choose how to handle violative material by providing "all information needed to remove harmful content on your [the social media platform's] own."[75] Crisp Thinking, another risk intelligence service, also submits

[72] Kate Crawford & Tarleton Gillespie, *What Is a Flag For? Social Media Reporting Tools and the Vocabulary of Complaint*, 18 NEW MEDIA & SOC'Y 410, 411 (2016).

[73] Klonick, *supra* note 10, at 1638.

[74] *The ActiveFence Story*, ACTIVEFENCE, https://www.activefence.com/about/ (last visited Jan. 18, 2022).

[75] *Harmful Content Detection*, ACTIVEFENCE, https://www.activefence.com/solutions/harmful-content-detection/ (last visited Jan. 18, 2022).

content that threatens to put a platform at risk using "existing T&S/trusted flagger workflows."[76]

As alluded to above on Crisp Thinking's website, "trusted flagger" programs, or programs that grant priority or expedited reviews for flags by certain actors, have also become an important tool for outside experts and civil society to help shape the enforcement of content.[77] Flagging is "a powerful rhetorical legitimation for sites . . . as they can claim to be curating on behalf of their user community and its expressed wishes,"[78] a reality that is complicated when other parties, and their accompanying wishes, are involved in flagging processes. YouTube's Trusted Flagger program, for example, includes a bulk-flagging option, prioritized review, discussion and feedback with YouTube, and online trainings (for some Trusted Flagger types only).[79] YouTube's initiative is open to individuals, government agencies,[80] and civil society organizations.[81] It appears, as mentioned above in Crisp Thinking's website, outside experts also use this trusted flagger status (and companies do not report it). Like concerns raised about the authenticity of civil society advisory boards, these trusted flagger statuses, priority flagging programs, and special reporting workflows may also face legitimacy challenges. Ultimately, companies choose who they would like to receive urgent flags from and, at least in the case of YouTube's program, participants must sign a non-disclosure agreement.[82]

While outside experts and civil society groups have wielded flagging to help shape the enforcement of content, flagging is not the only tool available to encourage changes to enforcement. At the heart of content

[76] *Platform Risk Intelligence*, CRISP THINKING, https://www.crispthinking.com/solutions/platform-risk-intelligence (last visited Jan. 20, 2022).

[77] YouTube refers to its program as the Trusted Flagger program, though other companies use priority flagging and review or dedicated reporting workflows for non-user flagger types.

[78] Crawford & Gillespie, *supra* note 72, at 412.

[79] *YouTube Trusted Flagger Program*, YOUTUBE, https://support.google.com/youtube/answer/7554338?hl=en (last visited Jan. 16, 2022).

[80] Internet Referral Units, law enforcement agencies with teams dedicated to reporting content on social media platforms that violate a platform's terms of service rather than local laws, have used this program as well. *See* Brian Chang, *From Internet Referral Units to International Agreements: Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114, 122 (2018) (discussing the presence of "special dedicated reporting channels" and "trusted flagger" status); Rabea Eghbariah & Amre Metwally, *Informal Governance: Internet Referral Units and the Rise of State Interpretation of Terms of Service*, 23 YALE J.L. & TECH. 542 (2021).

[81] *YouTube Trusted Flagger Program*, *supra* note 79.

[82] *Id.*

moderation is an exercise of interpretation of policies, an exercise that companies often expect content moderators to do in a matter of seconds. This enforcement interpretation, however, is also subject to influence through use of certain governance mechanisms such as datasets and databases and outsourcing enforcement to third-party companies.[83]

Enforcing content, and the associated interpretative process, have traditionally been confined to content *on* a given platform. External databases, compiled by outside organizations that use their own barometers of what may be objectionable, offensive, disturbing, or illegal. include images, text, videos, and other media across websites on the internet. Outside experts have turned increasingly to the compilation of these databases, selling access to companies to aid them in finding similar, or identical, content on their platforms. Organizations that focus on hate speech, terrorism, and extremism, for example, tout their databases as part of sales pitches on their websites. ActiveFence references its "Database of Evil"[84] and SITE Intelligence Group also sells its database as a service. SITE boasts that its database is the "largest commercially available global data set of confirmed terrorist and violent extremist online content." [85] Despite this proclamation, it says very little as to what "confirmed" means, and what sources or legal guidance was used to determine that an entity, and thus its content, is a terrorist actor. It is this type of "definitional ambiguity"[86] where alternative, or widened, interpretation and enforcement can take place. Outside actors that make their own policy determinations or use interpretative sources beyond just a platform's own Terms of Service, then could collect content that is far more expansive than a company's enforcement guidelines on a given topic. This expansion then invites more opportunities for erroneous enforcement (per a platform's policies) or opens the policies and rules for iterative revision,[87] one based on opaque practices, standards, and determinations by outside actors. This

[83] While social media platforms routinely use contractors to perform moderation work, that is not what is meant here. Though contractors are paid and managed by separate companies, these employees are still deeply intertwined with social media firms. What I mean by "outsourcing enforcement" is the use of outside experts in enforcement work to determine and report potentially violative content, through the use of enforcement detection algorithms, external content moderation, or external databases.

[84] *Harmful Content Detection*, *supra* note 75.

[85] *What We Do*, SITE INTEL. GRP. ENTER., https://ent.siteintelgroup.com/our-services.html (last visited Apr. 3, 2022).

[86] *See* Danielle Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep,* 93 NOTRE DAME L. REV. 1035, 1051 (2018).

[87] *See* Klonick, *supra* note 10, at 1648.

outsourced moderation may lead to even more aggressive enforcement when a third party service moderates on behalf of a social media platform, leading to further infringements on a user's online speech.

Other outside organizations, such as Two Hat, develop detection models using training data that companies can then purchase and deploy on their own content. Even this form of outsourcing is concerning because the development of algorithmic detection requires the use of massive training data to "teach" an algorithm to find content.[88] Platforms developing algorithms "in-house" source training data from their own corpora of content. Two Hat, for example, tells potential clients that its content moderation platform uses a "massive and diverse dataset" that will help customers reap benefits from other networks.[89] While harm undoubtedly transcends any one platform, outside of certain topics such as child sexual abuse material, consensus on what constitutes harm, and how much to allow, have traditionally been seen as the remit of individual platforms to decide. Proponents of these outsourced algorithms may argue that companies still retain ultimate authority through their policies and enforcement teams. This control, of course, is not threatened by an algorithm equipped with the task of finding content for a firm's moderators to review. However, these outsourced algorithms, and the third parties that develop them, assume that platforms view all "harms" equally and that there is no danger in creating tools shaped by training sets that do not closely mirror a platform's corpus of violative content. In doing so, imprecise algorithms can surface too much content, not only creating operational headaches but also opportunities for more reviewer error and, with time, demands on policy teams for clarity on enforcement guidelines. In addition, this imprecision has deleterious effects on users, particularly when the contested content involves politically charged notions of "extremism" which can stifle discussion and expression related to political dissent.

D. *One Set of Rules?*

[88] *See* Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance*, 7 BIG DATA & SOC'Y, at 3-5 (2020), https://journals.sagepub.com/doi/full/10.1177/2053951719897945 (offering a primer on the basics of the techniques of algorithmic moderation and the desired goals—whether to classify or match potentially violative material).

[89] *See Content Moderation Platform*, TWO HAT, https://www.twohat.com/solutions/content-moderation-platform/ (last visited Jan. 19, 2022).

Part I.E posited that while a single platform may act as a foreground decision maker, it is important to situate the other social media companies as experts in the background. One important goal behind this governance structure could be to ensure standardization in content policies, or a centralized form of speech and speech control on the internet. Standardization may also aid in deflection of regulatory scrutiny and public outcry.[90] One tactic to achieve this standardization is reflected through ad-hoc policy choices, such as behind the flurry of decisions to remove Alex Jones or de-platform Donald Trump.[91] Another tactic, however, is evidenced through the rise of industry-funded initiatives. For example, companies from Google to Twitter and Meta to Discord, to recognize the criticality of Trust & Safety work formed the Digital Trust & Safety Partnership (DTSP).[92] The organization seeks to develop what it sees as the best practices for the nascent Trust & Safety industry.[93] Even if these "best practices" are not binding, what might emerge from this industry initiative? At the bottom of a list of frequently asked questions is the question "Are all companies going to have the same policies now?"[94] The DTSP addresses an implicit anxiety behind these types of collaborations, writing that:

> No. It's important to note that this is not about individual company policies or decisions, but rather about how to do and assess the work of Trust and Safety overall. During the internal review process, each participating company will assess how it adheres to the commitments of trust and safety. While we are sharing insights, every platform continues to develop its own policies with respect to their own service.[95]

[90] *See infra* Part I.E.

[91] *See infra* Part I.E.

[92] DIGIT. TRUST & SAFETY P'SHIP, *supra* note 49.

[93] If this may sound similar to the work that other organizations like the Integrity Institute and the Trust & Safety Professionals Association do, you are not mistaken. However, the former is—I argue—not to be considered an expert actor because it is ultimately created and funded by the industry; instead it serves as a governance technique in furtherance of the goal, intentional or inevitable, of standardized sets of rules and practices. While the latter does seek to create recommendations and practices, I argue that this behavior is better reflective of creating new norms rather than an attempt to standardize policy and enforcement across companies. In other words, this last category is unique and exclusive to the technology companies as expert body.

[94] DIGIT. TRUST & SAFETY P'SHIP, *supra* note 49.

[95] *Id.*

However, there are other collaborations that suggest at least some degree of overlap for content policies. evelyn douek writes that industry collaboration to tackle child sexual abuse material (CSAM) spurred the collaboration of platforms identifying and reporting known CSAM content to databases run by certain child safety organizations.[96] The alignment in CSAM-related content policies is unsurprising: there's strong consensus around the abhorrent nature of this material and the category is relatively narrow.[97] Standardization, in other words, is to be expected here.

But what of other policy areas, where competing definitions and company postures abound? The industry's collaboration to create the Global Internet Forum to Combat Terrorism (GIFCT) helps to illustrate closer policy alignment. The GIFCT was created by Microsoft, Facebook, Google, and YouTube, and the organization maintains a database of content that ran afoul of platforms' terrorism policies.[98] There are now thirteen companies that have access to this database, meaning they can submit content they have removed *and* use the database to find uploads of content that other companies removed on their own platforms.[99] Until recently, the database only accepted submissions from companies using a narrowly defined taxonomy: terrorist actors on the United Nations Security Council's Consolidated Sanctions List.[100] If an actor in the content belonged to an entity on the Security Council list, then the content was also labelled based on certain themes: Imminent Credible Threat, Graphic Violence Against Defenseless People, Glorification of Terrorist Acts, and Recruitment and Instruction.[101] Though the actor list was narrowly scoped to avoid more politically-charged government terrorism lists, the vagueness

[96] evelyn douek, *The Rise of Content Cartels*, KNIGHT FIRST AMEND. INST. 8 (Mar. 23, 2021), https://s3.amazonaws.com/kfai-documents/documents/704838d2ec/3.23.2021_Douek_MW--To-Print-.pdf.

[97] *Id.*

[98] *Explainers*, GIFCT, https://gifct.org/explainers/ (last visited Jan. 21, 2022). There have been instances where the GIFCT database has taken content from actors outside of the technology platforms. *See* Metwally, *supra* note 21, at 45 ("[T]he GIFCT announced a 12-month pilot program to expand sharing URLs of known terrorist material to platforms. The GIFCT noted that it partnered with SITE Intelligence to collect 24,000 URLs, with the 'majority of new URLs shared amongst GIFCT member companies' coming from SITE Intelligence.").

[99] *Id.*

[100] *Broadening the GIFCT Hash-Sharing Database Taxonomy: An Assessment and Recommended Next Steps*, GLOB. INTERNET F. TO COMBAT TERRORISM 10 (2021), https://gifct.org/wp-content/uploads/2021/07/GIFCT-TaxonomyReport-2021.pdf [hereinafter GIFCT TAXONOMY REPORT].

[101] *Id.*

in the GIFCT taxonomy invites concern—and influence by other companies that participate in this database. For example, the GIFCT defines the Glorification of Terrorist Acts as "content that glorifies, praises, condones, or celebrates attacks after the fact."[102] If two companies on the list interpret "glorifies" or "condones" more aggressively than, say, other participating companies, the content is still added and distributed to the network. These definitional ambiguities force companies to think and see content more alike: an assessment of the GIFCT database advises that GIFCT members "increase the standardization of their terms of service."[103]

The GIFCT taxonomy has expanded to include a Content Incident Protocol (CIP) where participating companies share material of violent attackers' attack videos,[104] the first instance of which was the New Zealand terrorist attack in March 2019. The GIFCT provides no guidance as to the procedural requirements and considerations that must be made before companies agree that an attack and attacker fall under its CIP, marking the first time the database expanded beyond actors on the UN Security Council list. This expansion continues with the GIFCT considering and proposing new content categories such as "Promotion of terrorist or extremist ideologies."[105] Though the GIFCT maintains that each company has its own policies and enforcement decisions, an expansion of who and what can be added to and disseminated from the database creates tensions in the moderation enterprise. Choices from companies that interpret broadly and aggressively will exert pressure on others, resource-constrained smaller companies and platforms that do not have "in-house expertise" on this topic, especially at a time when more governments praise the use of the GIFCT database in addressing online violent extremism. This expansion creep continues to grow outside of the GIFCT as well, with calls for similar industry efforts to address synthetic media, disinformation, and even hate speech.[106] This expansion and standardization threatens the autonomy of platforms to devise their own philosophies around content moderation and prohibited speech, subjecting users to a set of norms and rules that extend beyond a single app or website with fewer avenues to express themselves fully.

[102] *Id.*

[103] *Id.* at 37.

[104] *Id.* at 10.

[105] *Id.* at 37.

[106] *See* douek, *supra* note 96, at 12.

Conclusion

These expert spheres produce winners and losers —some people, with their projects, that are heard loudly and repeatedly over others. These spheres are also increasingly intertwined with each other. For example, as discussed above, the GIFCT has been exploring an expanded taxonomy of content that companies would submit. In an analysis on the types of terrorism designation lists that should be used, the authors note that only using the UN list is what they term a "Limited Designation List."[107] A second option combines different governmental lists, an approach referred to as "Broader but Select Designation," while the third approach, "Expanded Designation," takes these two categories and adds to it lists maintained by companies and civil society groups.[108] The report recommends the GIFCT adopt the third option and work with the member companies and civil society to create an expanded list of terrorist actors and individuals.[109]

This paper illustrates the criticality of voices, perspectives, and ideologies that can shape the contours of permissible speech on social media platforms. It does not suggest that diversity of thought or a plurality of voices are vices. In fact, it is of the utmost importance. While this pluralism is immensely important, we must remain critical of whose voices are present and whose are not—and why and how they are excluded. Additionally, companies must be held accountable and disclose the names,[110] payments (if any), and work deliverables that come from any formal or informal consultation between expert actors. Companies could also publish calls that allow actors to apply and be considered; while companies retain the power to choose amongst submissions, a more transparent model introduces accountability if certain voices are repeatedly blocked from participation. While desirable in some ways and not in others, more information about the backgrounds of the inside experts (the company employees writing these policies) would also create a fuller picture of the different actors at play. At the very least, companies must consider more fully the range of experiences and ideologies available as they assess candidates for policy positions in their companies.

[107] GIFCT TAXONOMY REPORT, *supra* note 100, at 37.

[108] *Id.*

[109] *Id.* at 38.

[110] If disclosing names of individuals or organizations would pose security and safety risks, then at the very least a descriptor of the entity may suffice.

This entanglement of inside experts, outside experts, Trust & Safety organizations, industry collaboration initiatives, and the companies is poised to be even more complicated: expert engagement is increasingly mandated by law. As discussed in the introduction, the UK Online Harms regulatory approach mandates collaboration among companies and the sharing of knowledge and expertise. The European Union's Digital Services Act (DSA), among other requirements, mandates the use of trusted flaggers in certain circumstances.[111] The DSA empowers each member state's Digital Services Coordinator to award the status to an entity if it has particular expertise.[112] The EU approach also requires platforms to conduct risk assessments and design their mitigation measures involving groups that may be impacted by a platform's services and through the engagement with outside experts and civil society.[113] Experts and expertise are now the norm in our collective discussions on platform governance—more scrutiny on the expert bodies, their agendas, and the techniques they use to influence and control social networking platforms are needed now more than ever.

[111] *Proposal for a Regulation on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC*, at 55, COM (2020) 825 final.

[112] *Id.* at 55.

[113] *Id.* at 32.