Yale Journal of Law & Technology

Volume 23, Spring 2021

INFORMAL GOVERNANCE: INTERNET REFERRAL UNITS AND THE RISE OF STATE INTERPRETATION OF TERMS OF SERVICE

Rabea Eghbariah* & Amre Metwally**

The legal framework governing online speech relies on a distinction between the public and private sphere. A direct consequence of this distinction is the bifurcation between user and citizen. While the former is largely governed by private contractual norms—like a platform's terms of service—the latter is traditionally governed by public law norms. Governments, however, increasingly exploit this distinction and treat citizens as users: by engaging with the interpretation of private companies' self-regulation policies, governments are circumventing public law norms and fostering a new system of informal governance.

This article suggests the term informal governance to capture the nonbinding and opaque interplay between state actors and private content intermediaries, taking place in the shadow of the law and affecting online content moderation. Informal governance rests on the border of the public/private legal infrastructure and facilitates the circumvention of public law constraints. A distinctive feature of informal governance involves

* S.J.D. Candidate at Harvard Law School and human rights attorney at Adalah – The Legal Center for Arab Minority Rights in Israel. The author argued on behalf of Adalah in the case of *Adalah v. The Cyber Unit* discussed in this article. For insightful discussions that informed this article, I would like to thank Yochai Benkler, Noah Feldman, Hassan Jabareen, Fady Khoury, and Maroussia Lévesque. We would also like to thank the organizers and participants of the Cornell Law School Inter-University Graduate Conference. Finally, we would like to extend our thanks to Ben Rashkovich, Rafael Bezerra Nunes, and the rest of the team at the Yale Journal of Law and Technology for their incisive feedback and patience throughout the editing process. All errors are entirely our own.

** J.D. Candidate, Harvard Law School. The author was previously a Policy and Enforcement Manager (Political Extremism and Graphic Violence) at YouTube.

state institutions that subject their action to a private governance apparatus of a market player and engage with it to achieve their interests. Whereas informal governance is a conceptual framework, Internet Referral Units (IRUs) are its device in the content moderation enterprise.

IRUs are governmental units that submit non-binding requests to private content intermediaries, asking them to voluntarily remove content from their platforms based on an alleged violation of the platforms' own terms of service. In the past few years, governments have submitted, through IRUs, hundreds of thousands of voluntary takedown requests to online intermediaries, ranging from particular posts to entire websites, accounts or pages. But IRUs are not only facilitating the takedown of content: these units are also shaping the interpretation of companies' terms of service. This article assesses, problematizes, and contests the rise of informal governance by scrutinizing the understudied institution of IRUs.

Table of Contents

INTRODUCTION

Taking a look at the ease of communication in the age of social media might be deceiving. While it appears that the private is becoming more public—undermining spatial barriers and bearing emancipatory promises—public governance has simultaneously become more private, turning *citizens* into *users* and threatening to replace constitutional norms with terms of service. This metamorphosis of citizen to user signals shifting responsibilities from governments to private companies. While online speech is indeed subject to platforms' terms of service, social media companies are not alone in interpreting their own rules. In between these public and private spaces, a new system of *informal governance* of online speech is thriving.

States are entering the content moderation enterprise through the backdoor. A growing number of government-run units known as Internet Referral Units (IRUs)—embedded in law enforcement agencies—submit informal and legally non-binding requests to online intermediaries, asking them to takedown content that allegedly violates their own terms of service. Traditionally, governments request content removal based on violation of domestic laws, submitting binding court orders to the platforms. Now, governmental units circumvent this formal structure, scouring platforms for purported terms of service violations and submitting informal requests to the companies for review with little to no transparency or oversight. IRUs thus create an avenue for governments to request the global takedown of content that they possibly could not otherwise restrict.

Hundreds of thousands of takedown requests issued by IRUs have gone completely unchallenged in the past few years. What had started in 2010 as a law enforcement "Counter Terrorism" unit in the UK, referring content deemed "terrorist" to intermediaries for

"voluntary" takedown, spread ever since to many other countries in Europe and beyond.[1] While IRUs may be located within the confines of territorial borders, their reach extends far beyond any one country since they can submit referrals for extraterritorial content which would result in a global removal. The remit of these units has also expanded beyond flagging ill-defined terrorist content: IRUs also wield the power of the state to shape the interpretation of the companies' terms of service.

This article introduces informal governance as a conceptual framework that captures the opaque interplay between states and intermediaries affecting online speech, norms, and decision-making processes. A distinctive feature of informal governance is a systematized involvement of state institutions with the private governance apparatus of a market player, interpreting and enforcing its contractual terms of service based on ostensibly voluntary, nonbinding terms. The article conceptualizes, assesses, and challenges this phenomenon, recognizable through IRUs. Part I describes the legal infrastructure that gives rise to informal governance and situates it within a wider context of public-private partnerships and collaborations detectable in the digital age. Part II provides an overview of IRUs' structure and operation as well as an in-depth description of these units and their activity in five major jurisdictions. Part III turns to identify and assess the risks and implications of informal governance. Part IV contests the legitimacy of informal governance and argues that states should not be able to circumvent public law constraints by turning to informality and utilizing IRUs.

[1] For a comparative mapping of major IRUs, *see infra* Part II.B.

I. INFORMAL GOVERNANCE: PUBLIC, PRIVATE, AND THE RISE OF INFORMALITY

The "modern public squares"[2] are privately owned. Social media platforms, once celebrated for their democratizing effects,[3] are now simultaneously viewed as a threat to democracy itself. The largescale spread of social media has already spurred robust scholarly interest, introducing the "new-school" speech regulation,[4] declaring the rise of the "new governors,"[5] and warning of a "censorship creep."[6] Private control over public speech, however, is not new. The growing shift towards privately owned public squares had long ago posed some serious challenges to the legal norms governing these liminal public/private spaces[7] and led to efforts

[2] Packingham v. N.C., 137 S. Ct. 1730, 1737 (2017) (identifying social media platforms as the "modern public square" and holding a North Carolina statute barring registered sex offenders from using these platforms unconstitutional under the First Amendment).

[3] *See, e.g.*, Zeynep Tufekci, TWITTER AND TEAR GAS: THE POWER AND FRAGILITY OF NETWORKED PROTEST XXII (2017) (detailing the early praise given to social media platforms for their roles in helping to disseminate information and mobilize protestors during the 2011 Egyptian revolution).

[4] Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2298 (2014) (describing new school censorship as a move away from regulating speakers and "predigital" technologies to control of content intermediaries and platforms).

[5] Kate Klonick, *The New Governors: The People, Rules, and Processes Governing Online Speech*, 131 HARV. L. REV. 1598, 1603 (2018) (arguing that social media platforms are "New Governors" of speech that are "part of a new triadic model of speech that sits between the state and speakers-publishers. They are private, self-regulating entities that are economically and normatively motivated to reflect the democratic culture and free speech expectations of their users.").

[6] Danielle Citron, *Extremist Speech, Compelled Conformity, and Censorship Creep*, 93 NOTRE DAME L. REV. 1035, 1051 (2018) (defining censorship creep as "the expansion of speech policies beyond their original goals").

[7] Owen Fiss's work on free speech and social structure in the 1980s noted, for example, that "a shift from the street corner to CBS compels us to recognize the hybrid character of major social institutions; it begins to break down some of the dichotomies between public and private presupposed by classical liberalism." Owen M. Fiss, *Free Speech and Social Structure*, 71 IOWA. L. REV. 1405, 1414 (1986).

attempting to reconcile private spheres with public legal norms.[8] The inconsistent outcomes of the U.S. Supreme Court rulings on the applicability of the First Amendment on private property during the 1970s are perhaps the most illustrative of this struggle stemming from what may be viewed as the artificial construction of the public/private distinction in the context of speech regulation.[9]

These dilemmas have become more salient in the age of social media.[10] A developing line of cases pertaining to online speech suggests that a possible direction to resolve the tension between social media's public and private characters may reside in applying public law norms to speech hosted on these platforms. In *Packingham v. North Carolina*, the Supreme Court invalidated North Carolina's statute barring sex offenders from using social media platforms while identifying these platforms as the "modern public square."[11] Drawing on *Packingham*'s logic, courts have started to revisit the public forum doctrine[12] with regards to certain social media content. In *Davison v. Randall*, the Fourth Circuit invoked a public forum analysis and held unconstitutional blocking

---

[8] *See, e.g.*, Erwin Chemerinsky, *Rethinking State Action*, 80 NW. U.L. REV. 503, 505 (1985) (contesting the theoretical premises of the state action doctrine, suggesting a shift towards merit-based assessment of rights by the court regardless of the lack of state action); Gillian E. Metzger, *Privatization as Delegation*, 103 COLUM. L. REV. 1367, 1461-68 (2003) (proposing a "private delegation doctrine" that reformulates the state action doctrine in terms of delegation of government authority to private entities and asks whether the delegation is adequately structured to enforce constitutional constraints); Jonathan Peters, *The "Sovereigns of Cyberspace" and State Action: The First Amendment's Application (or Lack Thereof) to Third-Party Platforms*, 32 BERKELEY TECH. L. J. 989 (2017); *see also Developments in the Law State Action and the Public/Private Distinction*, 123 HARV. L. REV. 1248 (2010).

[9] For a detailed analysis of these decisions in the context of online speech regulation, *see* Klonick, *supra* note 5, at 1610-18.

[10] Jack M. Balkin, *Digital Speech and Democratic Culture: A Theory of Freedom of Expression for the Information Society*, 79 N.Y.U. L. REV. 1, 1-6 (2004) (suggesting that the digital age makes certain characteristics and conflicts more salient, rather than new).

[11] *Packingham*, *supra* note 2, at 1737.

[12] *See, e.g.*, Marsh v. Alabama 326 U.S. 501, 509 (1946) (holding that a company town would be treated as a government-owned public forum and is subject to the First Amendment).

a citizen from commenting on a government-run Facebook page.[13] Similarly, in *Knight First Amendment Institute v. Trump,* the Second Circuit affirmed the District Court for the Southern District of New York's holding that the "interactive space" of President Trump's originally private Twitter account (@realDonaldTrump) amounts to a public forum under the First Amendment.[14]

While these cases open the door for invoking the public forum doctrine, it is still important to note that courts have only applied this doctrine when a government-associated activity was at stake. The abovementioned cases leave open the question regarding the scope of the doctrine's application and whether social media companies themselves may be subject at all to legal burdens stemming from the public forum analysis.[15] The concurrence in *Davison v. Randall* makes this tension clear: "the interplay between private companies hosting social media sites and government actors managing those sites necessarily blurs the line regarding which party is responsible for burdens placed on a participant's speech."[16]

More recent developments, however, indicate that courts are not inclined to endorse claims that consider dominant social media companies as state actors. The Supreme Court recently discussed this issue with private entities more broadly, noting that "when a private entity provides a forum for speech, the private entity is not ordinarily constrained by the First Amendment because the private entity is not a state actor. The private entity may thus exercise

---

[13] *See* Davison v. Randall, 912 F.3d 666, 688 (4th Cir. 2019).
[14] Knight First Amend. Inst. v. Trump, 302 F.Supp.3d 541, 575 (S.D.N.Y. 2018), *aff'd*, Knight First Amend. Inst. v. Trump, 928 F.3d 226, 235 (2d Cir. 2019).
[15] *Compare* Jed Rubenfeld, *Are Facebook and Google State Actors?*, LAWFARE BLOG (Nov. 4, 2019), https://www.lawfareblog.com/are-facebook-and-google-state-actors, (analyzing precedents that might support classification of technology companies as state actors) *with* Alan Z. Rozenshtein, *No, Facebook and Google Are Not State Actors*, LAWFARE BLOG (Nov. 12, 2019), https://www.lawfareblog.com/no-facebook-and-google-are-not-state-actors. (arguing that Rubenfeld overstates the applicability of those precedents).
[16] *Randall*, 912 F.3d at 693 (Keenan, J., concurring).

editorial discretion over the speech and speakers in the forum."[17] Shaped in part by the Supreme Court's logic, just last year courts began weighing in on the question of whether social media companies constitute state actors that provide public fora. In a suit brought by the conservative organization Prager University against YouTube for allegedly censoring their videos and infringing their First Amendment rights by imposing age and other restrictions on some of the channel's videos, the court dismissed the claim that YouTube amounts to a public forum that is bound by the First Amendment. The Ninth Circuit held that "[d]espite YouTube's ubiquity and its role as a public-facing platform, it remains a private forum, not a public forum subject to judicial scrutiny under the First Amendment."[18] Furthermore, "[t]o characterize YouTube as a public forum would be a paradigm shift."[19]

The judgement in the *Prager University* case confirms that applying the public forum doctrine directly to social media companies seems unlikely in the near future. Furthermore, the desirability of this move is still normatively contested, with some commentators suggesting that applying First Amendment constraints to online content moderation by companies may result in an "internet nobody wants."[20] Despite growing efforts to regulate

[17] Manhattan Cmty. Access Corp. v. Halleck, 139 S. Ct. 1921, 1930 (2019).
[18] Prager Univ. v. Google LLC, 951 F.3d 991, 995 (9th Cir. 2020).
[19] *Id.* at 998.
[20] Klonick, *supra* note 5, at 1658-59 (arguing that interpreting the state action doctrine to apply to online platforms "would not only explicitly conflict with the purposes of §230, but would also likely create an internet nobody wants"); *see also* Enrique Amrijo, *Government-Provided Internet Access: Terms of Service as Speech Rules*, 41 FORDHAM URBAN L. J. 1499 (2015); Jack M. Balkin, *Free Speech Is a Triangle*, 118 COLUM. L. REV. 2011, 2025 (2018); Jonathan Peter, *The "Sovereigns of Cyberspace" and State Action: The First Amendment's Application-Or Lack Thereof-To Third Party Platforms*, 32 BERKELEY TECH. L. J. 989, 992 (2018).

and "break up" big tech companies,[21] the public/private distinction is, seemingly, here to stay, and it continues to shape the legal framework of contemporary online content regulation.

This distinction makes it possible to talk about "private law," in opposition to "public law," and grounds the admittedly contested idea that a private company is inherently different from a government regardless of its centrality in facilitating or creating a "public" sphere. Online content intermediaries and governments, therefore, function in different legal realms and are governed by different legal norms. A direct consequence of this distinction is the bifurcation between *user* and *citizen*. While the former is largely governed by private contractual norms—like a platform's terms of service—the latter is governed by public law norms.

At the same time, however, this distinction between public and private law as they apply to governments and online intermediaries, respectively, does not mean that these institutions are parallel entities operating in completely independent realms, with the private realm free from state power. The public and private spheres are concepts created by law, which in turn defines the different *forms* in which state power is exercised in those realms.[22]

---

[21] *See, e.g.*, Marcy Gordon, *Democrats call for Congress to rein in, break up Big Tech*, ASSOCIATED PRESS (Oct. 6, 2020), https://apnews.com/article/technology-50e69e921c6699a3edbd730c12292436; Makena Kelly, *Alexandria Ocasio-Cortez supports Big Tech breakup plan laid out by Elizabeth Warren*, VERGE (May 3, 2019, 11:46 AM), https://www.theverge.com/2019/5/3/18528234/alexandria-ocasio-cortez-big-tech-break-up-plan-elizabeth-warren-endorsement; Sheelah Kolhatkar, *How Elizabeth Warren Came Up with a Plan to Break Up Big Tech*, NEW YORKER (Aug. 20, 2019), https://www.newyorker.com/business/currency/how-elizabeth-warren-came-up-with-a-plan-to-break-up-big-tech.

[22] For an overview of the development of the public/private distinction in modern political and legal thought, see generally Morton J. Horwitz, *The History of the Public/Private Distinction*, 130 U. PA. L. REV. 1423 (1982). Both legal realism and legal feminism have extensively challenged the public/private distinction, exposing the central role of state power in the private sphere. *See, e.g.*, Morris R. Cohen, *Property and Sovereignty*, 13 CORNELL L. REV. 8 (1927); Robert Hale,

Governments, especially more powerful governments,[23] have the capacity to effectively impose regulation on online intermediaries, subject them to liability under domestic laws, or block access to their websites altogether. In this sense, governments possess powerful tools to incentivize companies to comply and cooperate with them. Governments can sustain a largely predictable legal environment for companies to operate in; alter the atmosphere for platforms by imposing liability, regulation, or taxation (where jurisdictionally applicable);[24] or simply threaten to do so. In contrast, companies can either respond by complying to government's requests to ensure a more predictable and comfortable legal landscape; refuse to cooperate at the expense of being exposed to potential costs, regulation, or blocking procedures; or "exit" the country completely, although the latter is highly unlikely. This set of asymmetric bargaining powers, taking place in the shadow of the law,[25] has the ability to neutralize significant resistance from these companies, although it is not completely absent.[26]

Coercion and Distribution in a Supposedly Non-Coercive State, 38 POL. SC. Q. 470 (1923); Susan S. Boyd, "Challenging the Public/Private Divide: An Overview," *in Challenging the Public/Private Divide: Feminism, Law and Public Policy* 3 (1997).

[23] As discussed later, we acknowledge not all governments have the same capacity to reach, liaise, influence, or communicate with these intermediaries. Countries from the Global North typically possess higher influence vis-à-vis social media companies.

[24] For example, the Australian Government in the aftermath of the Christchurch shootings passed the Sharing of Abhorrent Violent Material Act which opened the possibility of imposing criminal offenses on companies' executives for failing to quickly remove graphic material from social media platforms. Criminal Code Amendment (Sharing of Abhorrent Violent Material) Act 2019, https://www.legislation.gov.au/Details/C2019A00038. Other examples of regulation, or threats thereof, are demonstrated throughout the article.

[25] We use this term in reference to the work of Mnookin and Kornhauser. Robert H. Mnookin & Lewis Kornhauser, *Bargaining in the Shadow of the Law: The Case of Divorce*, 88 YALE L. J. 950 (1979).

[26] *Cf.* Alan Z. Rozenshtein, *Surveillance Intermediaries*, 70 STAN. L. REV. 99 (2018). For detailed information about companies' cooperation with IRUs, see Part II below. It shall be noted that users possess some bargaining power which may affect these dynamics, by protesting or threatening to exit the platform.

This legal infrastructure enables and fosters some practices that can hardly be classified as either completely public or utterly private. Governments are now increasingly exploiting the fact that private content intermediaries employ self-regulation over content on their platforms. From the platforms' point of view, self-regulation is necessary from a business perspective to attract users and advertisers, even when protected from liability under legal regimes such as Section 230 of the Communications Decency Act.[27] Given the companies' self-regulation mechanisms, governments have started designing regulatory schemes that rely on the companies' self-regulation apparatus, in a move towards regulating self-regulation. The most remarkable example for this move is Germany's 2017 Network Enforcement Act (NetzDG), which requires large social media companies, such as Facebook, Twitter, and YouTube, to make their own determinations on the legality of content according to German law and promptly remove "illegal content" or face a fine of up to 50 million euros.[28] Civil society organizations have widely criticized the law, and Germany's director at Human Rights Watch described the law as "vague, overbroad, and turns private companies into overzealous censors to avoid steep fines, leaving users with no judicial oversight or right to appeal."[29]

However, this bargaining power is much more dispersed among individual users in comparison to the concentrated power held by governments or companies. In the case of IRUs, the interplay between governments and companies is taking place behind closed doors, which renders users' capacity to participate in the conversation largely ineffective.

[27] Balkin, *supra* note 20, at 2022-23; *see also* Monika Bickert, "Defining the Boundaries of Free Speech on Social Media," *in The Free Speech Century* 254 (2019).

[28] Gesetz zur Verbesserung der Rechtsdurchsetzung in sozialen Netzwerken [NetzDG] [Network Enforcement Act], Sept. 1, 2017, BUNDESGESETZBLATT, Teil I [BGBL I] at 3352 (Ger.).

[29] *Germany: Flawed Social Media Law*, HUMAN RIGHTS WATCH (Feb. 14, 2018), https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law.

Imposing regulation on companies' content moderation mechanisms is not the only path that governments have taken to utilize the companies' self-regulation apparatus. A remarkable alternative has been exerting informal pressure to take down content and prevent its dissemination. By turning to informality, governments are ostensibly able to escape constitutional and administrative law constraints, address citizens as users, and urge intermediaries to take down content that would otherwise be illegal or too costly to sanction. It is in between these public and private domains that informality breeds, and gives birth to what may be termed informal governance.[30]

This article suggests the term informal governance to capture the new system of informal, nonbinding, and nontransparent interplay between state actors and private content intermediaries, taking place in the shadow of the law and affecting online speech. Informal governance includes but also goes beyond merely employing informal pressure to take down content: a distinctive feature involves state institutions that subject their action to a private governance apparatus of a market player and engage with it to achieve their interests. Although this engagement is often highly institutionalized, it is still informal in the sense that it is a combination of ostensibly non-coercive, largely non-transparent, and effectively unchecked engagements. Informal governance provides a conceptual framework that centers the background interactions between states and private platforms as an important

[30] As far as we are aware, the term informal governance has not been introduced in legal literature. Thomas Christiansen, Andreas Føllesdal and Simona Piattoni used the term to describe "the operation of informal networks which link policy makers to client groups" and "when participation in the decision-making process is not yet or cannot be codified and publicly enforced." Thomas Christiansen, Andreas Føllesdal & Simona Piattoni, *Informal Governance in the European Union: An Introduction*, *in* INFORMAL GOVERNANCE IN THE EUROPEAN UNION 1, 6 (2003). Although we use the term to describe informal governance of virtual speech, the term may be relevant to other legal fields.

arena shaping the norms and decision-making processes governing online content. It brings into clearer view the interplay between the "public" and the "private" which otherwise remains overshadowed by centering the "private" or "new" governors.[31]

Informal governance does not occur in a vacuum. The public-private cooperation has gained an infamous reputation in the digital context. At the beginning of the millennia, Michael Birnhack and Niva Elkin-Koren examined the post-9/11 legal environment pertaining to Internet Service Providers (ISPs) and pointed out that the market's "invisible hand"[32] had been hijacked and replaced by an "invisible handshake," namely, "the informal coordination between the government and market players, which is executed in a legal twilight zone."[33] Comparably, before the rise of speech regulation on social media, Seth F. Kreimer coined "censorship by proxy" to describe governments' pressure on intermediaries to "prevent Internet communications from reaching their intended audiences" by targeting "weak links" of the internet chain without the ability to "determine how dialogue has been deformed."[34]

---

[31] While the "new governors" are certainly foundational in facilitating, moderating, and "governing" the network—that is, generating, interpreting, and enforcing norms in the virtual sphere—they are not fully independent agents and their governance scheme remains amenable to informal, sometimes invisible, influences by state actors. The term "governance" used by Klonick draws from Rhodes to describe "the interplay between user and platform: a 'dynamic' and 'iterative' 'law making process'; 'norm generating' 'individuals'; and 'convergence of process and outcomes.'" Klonick, *supra* note 5, at 1617, quoting R. A. W. Rhodes, *The New Governance: Governing without Government*, 44 POL. STUD 652 (1996). We expand this understanding to include and emphasize the informal interplay between platforms and governments that affect users, rather than the direct interplay between a user and platform. *See also* Balkin, *supra* note 20, at 2021 (discussing the meaning of private governance).

[32] Michael D. Birnhack & Niva Elkin-Koren, *The Invisible Handshake: The Reemergence of the State in the Digital Environment*, 8 VA. J. L. & TECH. 6 (2003).

[33] Niva Elkin-Koren & Eldar Haber, *Governance by Proxy: Cyber Challenges to Civil Liberties*, 82 BROOK. L. REV. 105, 115 (2016).

[34] Seth F. Kreimer, *Censorship by Proxy: The First Amendment, Internet Intermediaries, and the Problem of the Weakest Link*, 155 U. PA. L. REV. 11, 17, 28 (2006).

Yochai Benkler has also identified an "extralegal public-private partnership" that allows governments to achieve "results that would have been practically impossible to achieve within the bounds of the Constitution and the requirements of legality," in his assessment of a sequence of informal pressure employed by officials on different private actors to prevent services from WikiLeaks.[35] Similarly, Derek Bambauer has described instances of what he calls "soft censorship," which includes "persuading intermediaries to restrict content" and argued that direct regulation is normatively preferable because it is overt.[36] More recently, Jack M. Balkin coined the term "new-school speech regulation" to describe speech regulation through a range of governmental practices that target private internet infrastructures, including social media companies, instead of regulating the speaker or the publisher directly. These practices include "public/private cooperation and co-optation," through which governments seek to "coax, cajole, or coerce" intermediaries, by offering "a combination of carrots and sticks, the most important being legal immunity."[37]

In this context, a new institution of informal governance has emerged. IRUs are government-run units that submit non-binding requests to private content intermediaries, asking them to "voluntarily" remove content from their platforms. Although the content may also be illegal according to domestic law, IRUs use informal channels and submit requests based on an alleged violation of the platforms' own terms of service. Government bodies traditionally submit court orders to social media platforms in instances where content violates local laws; if a platform finds that

---

[35] Yochai Benkler, *A Free Irresponsible Press: Wikileaks and the Battle over the Soul of the Networked Fourth Estate*, 46 HARV. CR.-CL. L. REV. 311, 314 (2011); *see also* Yochai Benkler, *WikiLeaks and the Protect-IP Act: A New Public-Private Threat to The Internet Commons*, 140 DAEDALUS 154 (2012).
[36] Derek Bambauer, *Orwell's Armchair*, 79 U. CHI. L.R. 863, 867 (2012).
[37] Balkin, *supra* note 4, at 2325.

the content falls short of its community guidelines but does violate a country's laws, then it will geo-block the material. IRUs, instead, focus on content that falls afoul of community standards, the result of which is a global removal, an effective method to circumvent jurisdictional limitations resulting from legal takedown requests. Although companies ostensibly have the discretion to review referred content, the "voluntary" nature should invite scrutiny as discussed later in this article.[38]

While literature documented and analyzed incidents and practices of informal public-private partnerships, the massive uptick in informal governance and the shift it is taking through the rise of IRUs has largely gone understudied.[39] IRUs are tools through which governments not only signal to companies the specific content they are interested in taking down, but also contribute to shaping the desirable interpretation of the companies' terms of service. Moreover, when social media companies are facing state requests to take down content, the companies are in fact making decisions at the behest of a "Repeat Player,"[40] namely, a party that gains advantage through engaging in many similar interactions. Engaging in repeated, informal requests through IRUs helps governments assert

---

[38] *See infra* Part III.A.2.

[39] A remarkable exception in the literature is 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, 122-23 (2018) (studying the legality of UK and EU IRUs in light of international human rights law). Some other works briefly discuss or mention IRUs, but do not particularly focus on IRUs or extensively analyze their activity. Works the authors are aware of are DAVID KAYE, SPEECH POLICE: THE GLOBAL STRUGGLE TO GOVERN THE INTERNET, 79–81 (2019); Citron, *supra* note 6, at 1043; Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27, 45 (2019); Daphne Keller, *Who Do You Sue?: State and Platform Hybrid Power Over Online Speech*, HOOVER INST. 7 (2019), https://www.hoover.org/sites/default/files/research/docs/who-do-you-sue-state-and-platform-hybrid-power-over-online-speech_0.pdf; Molly Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, 60 VA. J. INT'L. L. 363, 380 (2020).

[40] Marc Galanter, *Why the "Haves" Come Out Ahead: Speculations on the Limits of Legal Change*, 9 L. & SOC'Y REV. 95, 108 (1974).

extralegal pressure to achieve goals that might otherwise be illegal or too costly (politically or economically) to achieve.

What makes this model remarkable is not only that governmental units subject themselves to terms of service and act to enforce them, but also that it occurs through highly institutionalized informal interactions, operating continuously in the background to foster takedowns of online content ranging from particular posts to search results to entire websites, accounts and pages. It is important to note, however, that IRUs are not only becoming more popular, as witnessed by expanded remits, their legitimacy is also being affirmed by legal institutions as well. In *Adalah v. The Cyber Unit*, one of the first cases challenging the legality of an IRU worldwide, the Israeli Supreme Court recently found that the state did have the authority to operate a referral unit.[41] In addition to the growing popularity and legitimation of IRUs, traditional methods of informal governance, where state actors employ extralegal pressure to circumvent existing legal avenues, may still be practiced as well.

Before we delve into describing these institutions, it is productive to clarify that this article does not make an absolute normative claim against informal governance in all contexts. It does, however, identify informal governance as an organizing logic in the way that IRUs function and calls this logic into question. While some may contend that informal governance is simply a more efficient way to "do things" and enforce rules, this article brings into question the "how," "what," and "who" of "doing things." It scrutinizes IRUs as institutions that foster informal governance, analyzes their potential risks, and challenges their legitimacy.

[41] HCJ 7846/19 Adalah v. The Cyber Unit (2021) (Isr.). For discussion of the case, see *infra* Part IV.

II. INTERNET REFERRAL UNITS: ACTIVITY, CHARACTERISTICS, AND SCOPE

IRUs are now blooming with minimal transparency and almost no scrutiny over their activities. Instead of what is by now relatively familiar incidents of politicians employing informal pressure on content intermediaries, IRUs are institutionalized governmental units operating in liminal public/private legal spaces by issuing non-binding requests to online intermediaries. IRUs are now acknowledged and operating in several countries, including the UK, France, and Israel.[42] The EU IRU was also established under Europol in 2015, and the European Commission has been calling on EU members to supplement its efforts by establishing additional national IRUs.[43]

One commonality between the different IRUs arises from the fact that these governmental units outsource takedown decisions to private content intermediaries. They do so by "flagging" online content to intermediaries, urging platforms to "voluntarily" take down content, according to their own terms of service, rather than submitting binding orders according to state law. Additional commonalties include both an overwhelming focus on content categorized as "terrorist" or "inciting," and little to no transparency or public oversight mechanism. In addition to these shared general characteristics, each unit is designed and operates differently in each state or jurisdiction, despite the fact that takedown requests may pertain to content published worldwide. These differences include the degree of existing legal authorization; the types of content referred; the standard of review (terms of service violation only or

[42] *See infra* Part II.B.
[43] Chang, *supra* note 39, at 121.

in conjunction with domestic law violation); and the degree of transparency or oversight mechanisms available, if any.[44]

It would be an understatement, however, to characterize IRUs as entities that only submit requests to companies for voluntary removal; this article suggests viewing IRUs as forces that influence the companies' terms of service by wielding the power of a handful of governments to continuously engage in the *interpretation* of companies' terms of service. It is also important to note, in this context, that not all states have the same leverage to bargain with global private platforms. Some less powerful governments may even find it hard to contact the private companies and establish relationships that enable informal governance.[45] Furthermore, it is crucial to remember that global private platforms are not weak actors and are often incorporated and embedded in Western states in a way that may be amenable to reproducing global hierarchies. This may provide some explanation as to why IRUs are most detectable in states that self-identify as Western liberal democracies (although they are not necessarily limited to them by design). Additionally, whereas some regimes have legal cultures that regularly engage in blunt censorship, informal governance may provide an effective tool to engage in some censorial activity

---

[44] Some may argue that legal authorization to refer content to platforms, even if the decision remains at the hands of the companies, can be seen as a formal action. Alternatively, a certain degree of transparency or oversight (or institutionalization more generally), could qualify these interactions as formal. Although differences in degree can eventually become differences in kind, we discuss and characterize informal governance not only in terms of authorization or lack of transparency, but even more so as a non-binding, nontransparent, and unchecked engagement between "public" and "private" actors, taking place against the background of asymmetric bargaining powers, and influencing the norms and decision-making processes.

[45] Rebecca Hamilton, *Governing the Global Public Square*, HARV. INT'L. L. J. (forthcoming 2021) (challenging the traditional assumption that "a state can secure the attention and engagement of a platform" and describing the case of Sri Lanka as an example of the fact that "sometimes a government cannot even manage to contact those who work at a platform"). *See also* Chinmayi Arun, *Facebook's Faces*, 135 HARV. L. REV. FOR. (forthcoming 2021) (theorizing the varying degrees of influence of different states and publics on Facebook).

(especially under the pretext of national security) while simultaneously maintaining an appearance of overall neutrality with regards to content takedowns made by private intermediaries. As such, very little is known about IRUs, and it may also be the case that similar units or behavior is much more abundant across the globe, perhaps in a less institutionalized manner.

This section seeks to shed light on the largely unknown phenomenon of IRUs. It starts by examining the structure and operation of an IRU, discussing three main characteristics: (1) the voluntary mode of removal behind IRU referrals; (2) the transparency and oversight mechanisms available; and (3) the volume and type of content referred by IRUs. (Critiques of the structure and operation of these units are left to Part III.) Part II.B provides a comparative lens by highlighting the operation of these units in the United Kingdom, the European Union, France, Israel, as well as pointing out the attempts to establish a similar unit in the United States.

A. Structure and Operation of IRUs

Despite the proliferation of different modalities of IRUs, they appear to possess certain similarities in terms of their design, structure, and operations. This section identifies and delineates on three main characteristics of IRUs.

1. Voluntariness

IRUs issue requests to companies for "voluntary" removal. The unit's requests are based on the companies' own terms of service and are, therefore, formally non-binding and subject to the discretion of the companies. This voluntariness is explained, for example, in an EU IRU report: "a referral activity does not constitute an enforceable act," rather, the decision "is taken by the concerned service provider under their own responsibility and accountability

(in reference to their Terms and Conditions)."[46] These referrals are said to be built on "cooperation with the private sector" and "the trust-based relationship with the industry."[47]

How is this voluntariness manifested? Referral requests are not issued through traditional legal mechanisms, meaning they do not involve binding orders by judicial or semi-judicial bodies. Rather, these requests are issued by administrative agencies that utilize different communication channels with the companies, ranging from regular in-product flagging to special dedicated communication channels, and therefore subject the requests to the company's review processes.

Traditionally, if a government believes that a piece of content violates local law, its normal option is to submit a legal order from a judicial body to a company asking for its removal.[48] Companies, particularly those with extensive resources available to review each request, can then determine how best to proceed, ranging from blindly accepting a removal request to making its own determination after legal input that the content does not violate local law. Smaller companies are often at the highest risk of engaging in minimal review due to a lack of operational expertise or legal guidance.[49] IRUs, however, do not exhaust this traditional legal channel and turn, instead, to flag content directly to the company and ask for its removal based on an alleged violation of the companies' own terms of service. Large and small companies alike are now adjudicating government removal requests on more than

---

[46] EUROPOL, EU INTERNET REFERRAL UNIT TRANSPARENCY REPORT, 4 (2017), https://www.europol.europa.eu/publications-documents/eu-internet-referral-unit-transparency-report-2017 [hereinafter EU IRU TRANSPARENCY REPORT].
[47] *Id.*
[48] For a helpful breakdown of removal orders from government agencies, *see* Brian Fishman, *Crossroads: Counter-terrorism and the Internet*, 2 TEXAS NAT'L SEC. REV. 82, 90 (2019). Brian Fishman also serves as Facebook's Director of Counterterrorism and Dangerous Organizations.
[49] *Id.*

one front: traditional legal orders as well as voluntary referrals based on terms of service violations.

The voluntary requests—a product of an administrative rather than judicial process—are delivered to the companies either through direct communication with the company or by utilizing a more systematized channel. While in some cases IRUs may resort to specially designated channels to contact the companies,[50] in other cases they may use a platforms' already existing reporting mechanisms which are also accessible to users. Even when using the platforms' existing mechanisms, IRUs may still benefit from special status. YouTube's Trusted Flagger program provides a good illustration. A Trusted Flagger status is usually given to individuals or organizations with significant expertise in a particular area or those who have flagged content for removal with high accuracy (meaning content they report does, in fact, actually violate a company's terms of service). Private users, non-governmental organizations, and government agencies themselves are all invited to apply.[51] A Trusted Flagger is allowed, among other things, to flag content in bulk—whereas all other users can only flag one piece of content at a time—and the opportunity to engage in "ongoing discussion and feedback with YouTube about various content areas."[52] In addition to the opportunity to engage in a feedback loop with the company, Trusted Flaggers are given "[p]rioritized flag reviews for increased actionability" as well.[53] In other words, with conferring this designation, the company grants an IRU an option to

[50] Chang, *supra* note 39, at 122.
[51] YouTube, *YouTube Trusted Flagger program*, YOUTUBE HELP CENTER, https://support.google.com/youtube/answer/7554338?hl=en. For government agencies, there appears to be a restriction in applying to the program if the country has a "history of human rights abuses or a suppression of speech" then additional review by company staff may be required. *Id.*
[52] *Id.*
[53] *Id.*

engage in a special, prioritized, and expedited communication channel with regards to content takedowns.

2. Transparency & Oversight

There is a dearth of transparency reporting from IRUs about their referrals.[54] While some units have scant amounts of statistics included in publicly accessible reports,[55] the reporting often lacks meaningful depth to properly account for the extent, quality, and efficacy of IRU removal requests. These reports, when made available, do not appear to be a consistent practice across various IRUs, and the quality produced and detail provided vary significantly. As noted in a Global Network Initiative report "[e]ven those that have issued transparency reports, such as the French mechanism and the EIRU [Europol's IRU], tend to mostly include cumulative statistics on referrals or 'content removed' in ways that make verification and accountability a major challenge."[56]

Social media companies also fail to shed light on IRU behavior, despite many platforms producing reports on enforcement trends on their own websites. Companies often publish transparency reports, detailing removal volumes, takedown reasons, and other operational metrics, though the frequency of such reports varies by company. An examination of reports published by YouTube, Twitter, and Facebook shows that these companies do not segment

---

[54] Chang, *supra* note 39, at 145.
[55] *See infra* Part II.B.
[56] Jason Pielemeier & Chris Sheehy, *Understanding the Human Rights Risks Associated with Internet Referral Units*, MEDIUM (Feb. 25, 2019), https://medium.com/global-network-initiative-collection/understanding-the-human-rights-risks-associated-with-internet-referal-units-by-jason-pielemeier-b0b3feeb95c9.

the data to reflect referrals from IRUs.[57] Whereas YouTube and Twitter mention governmental takedown requests based on terms of service violations, Facebook's Transparency Report has no mention—whether explicit or implicit—of government flagging on the grounds of a purported terms of service violation (compared to legal processes).[58] However, even when YouTube and Twitter reports mention government requests to takedown content based on terms of service, the various transparency metrics fail to provide a realistic account of IRU involvement.[59]

The lack of transparency into IRUs' actions is not the only barrier to understanding their activities. Effective oversight mechanisms to assess the unit's actions as a government authority are also deeply lacking as well. While some IRUs are subject to parliamentary oversight or maintain a form of internal review mechanism, as detailed in Part II B below, the efficacy of these mechanisms remains heavily in question.[60] None of the IRUs

---

[57] *See generally Facebook Transparency Report*, FACEBOOK, https://transparency.facebook.com; *Twitter Transparency Center*, TWITTER, https://transparency.twitter.com/; *YouTube Community Guidelines Enforcement*, GOOGLE, https://transparencyreport.google.com/youtube-policy/removals?hl=en.

[58] *See Facebook Transparency Report*, *supra* note 57.

[59] *See Government Requests to Remove Content*, GOOGLE, https://transparencyreport.google.com/government-removals/overview?hl=en; *Government Terms of Service Reports Jul 1 – Dec 31, 2016*, TWITTER (Mar. 21, 2017), https://transparency.twitter.com/en/reports/rules-enforcement-archive/gtr-2016-jul-dec.html; *Government Terms of Service Reports Jan – June 30, 2017*, TWITTER (Sept. 19, 2017), https://transparency.twitter.com/en/reports/rules-enforcement-archive/gtr-2017-jan-jun.html; *Government TOS Reports Jul – Dec 31, 2017*, TWITTER (Apr. 5, 2018), https://transparency.twitter.com/en/reports/rules-enforcement-archive/gtr-2017-jul-dec.html.; *Legal Demands, in Removal Requests*, TWITTER, https://transparency.twitter.com/en/reports/removal-requests.html#2019-jul-dec.

[60] Civil society groups, including the Center for Democracy and Technology (CDT), the ACLU, and Access Now, have repeatedly criticized in the past few years the lack of transparency and oversight mechanisms with regards to IRUs activity. Similarly, examining the EU IRU's parliamentary oversight, Brian Chang concluded that "while the new Europol Regulation grounds the IRU within a legal framework, it does not address the significant due process and transparency concerns about the IRU. It does create political oversight in the form of a Joint Parliamentary Scrutiny Group, but this oversight will be hindered by the fact that

maintain a public mechanism of oversight which enables individuals or interested parties to call an IRU's specific removal request into question or refer the request for further review. The lack of effective oversight does not seem to be merely a deficiency in the operation of IRUs but rather an integral part of its design: the operational structure of these units is modeled around informal communications with private intermediaries which, as such, circumvent judicial review or formal processes that impose legal constraints including transparency and oversight.

3. Referred Content and Volume

The categories of content referred under IRUs' activities may vary, although most of the content referred is labelled "terrorist." In fact, the vast majority of IRUs were formed with the specific mandate of finding and reporting "terrorist" content online. Some units have seen their mandate grow to also include child sexual abuse material, "illegal immigration," and other crimes of concern for law enforcement agencies. Most recently, IRUs have been also used to report misinformation regarding the COVID-19 vaccines and ask for the takedown of Facebook groups that hosted such content.[61]

Finally, as both mature and nascent IRUs grow, referral volume has dramatically expanded over time. These staggering increases are noted in more detail in Part II B in a comparative lens

---

the European Parliament does not have any more privileged access to documents than an ordinary EU citizen. While it contains data protection safeguards, these are presently insufficient to address the freedom of expression and due process concerns raised by the EU IRU." Chang, *supra* note 39, at 190.

[61] *Israel Starts Covid Vaccine drive as Facebook Groups Taken Down*, GUARDIAN (Dec. 20, 2020), https://www.theguardian.com/world/2020/dec/20/facebook-takes-down-groups-spreading-lies-about-covid-vaccine-in-israel; Netael Bandel, *Facebook Removes Hebrew-language Groups Spreading False Coronavirus Vaccine Information*, HAARETZ (Dec. 20, 2020), https://www.haaretz.com/israel-news/tech-news/.premium-facebook-deletes-four-groups-posting-false-coronavirus-vaccine-information-1.9386408.

by country. It is important to note that the referral increases not only shows a rise in reported, and removed, content but also potentially much higher takedown volume than publicly available data may suggest. Each referral may include multiple pieces of content, each of which, in turn, can range from a single post to an entire page or website. Moreover, depending on a company's particular review workflow, a flagged piece of content could trigger review of a user's other content or a user's entire account. A moderator may disable the entire account, removing with it all of the content on the profile, bringing into question the true volume of removed content from IRU requests.

B. IRUs: A Comparative Lens

All known IRUs share a rhetoric of voluntary referrals, limited transparency and oversight, a predominant focus on referring "terrorist content," and notable increases in the volume of referred content. In the following section, we examine four particular IRUs—the United Kingdom's IRU (CTIRU), the European Union IRU, the French IRU (OCLCTIC), the Israeli IRU—and one country's opaque efforts to explore the establishment of a referral unit: the United States.

The focus on these particular IRUs stems from the fact that these units are the most prominent IRUs identifiable, both in terms of the volume of content referred and the relatively available information regarding their activity. Furthermore, the Israeli case invites special scrutiny: the Israeli IRU is the only one operating as part of a state apparatus that maintains decades-old occupation of a people that amounts to an apartheid regime.[62] The use of an IRU in

[62] The international community has recently started to reckon with long-standing claims by Palestinians describing Israel as a state manifestation of a settler-colonial project that practices apartheid. Most recently, Human Rights Watch

light of this reality, as discussed later, raises unique concerns as to the silencing of legitimate Palestinian speech and resistance under the pretext of anti-terrorism laws.[63]

Other countries have also acknowledged operating IRUs or a similar work model based on the voluntary removal by online intermediaries, although their scope of activity remains unclear.

published a thorough report finding that Israel is committing the crimes of apartheid and persecution under international law. *See* Human Rights Watch, A THRESHOLD CROSSED: ISRAELI AUTHORITIES AND THE CRIMES OF APARTHEID AND PERSECUTION 10 (2021), https://www.hrw.org/sites/default/files/media_2021/04/israel_palestine0421_web_0.pdf. For a review of Human Rights Watch's report in light of the Palestinian intellectual tradition that identifies Israel as a settler-colonial project, *see* Noura Erakat, *Beyond Discrimination: Apartheid is a Colonial Project and Zionism is a Form of Racism*, EJIL:*TALK!* (July 5, 2021), https://www.ejiltalk.org/beyond-discrimination-apartheid-is-a-colonial-project-and-zionism-is-a-form-of-racism/.

[63] *See generally* Amy Braunschweiger, *Witness: How Israel Muzzles Free Expression for Palestinians*, HUM. RTS. WATCH (Dec. 17, 2019), https://www.hrw.org/news/2019/12/17/witness-how-israel-muzzles-free-expression-palestinians#. The case of Palestinian poet Dareen Tatour who was sentenced for 5 months in prison for her poems attracted widespread attention in the past few years. *See, e.g.*, Gideon Levy & Alex Levac, *In 2016 Israel, A Palestinian Writer Is in Custody for Her Poetry*, HAARETZ (May 21, 2016), https://www.haaretz.com/israel-news/.premium-in-2016-israel-a-palestinian-writer-is-in-custody-for-her-poetry-1.5385083; Mustafa Abu Sneineh, *Israel Convicts Palestinian Poet Dareen Tatour of Facebook 'Incitement'*, MIDDLE EAST EYE (May 4, 2018), https://www.middleeasteye.net/news/israel-convicts-palestinian-poet-dareen-tatour-facebook-incitement.

These countries include Germany,[64] Spain,[65] Austria,[66] Belgium,[67] Italy,[68] the Netherlands,[69] and Switzerland.[70] It is noteworthy that the German case is particularly remarkable: the operation of an IRU

---

[64] Matthias Monroy, *German Police Launches "National Internet Referral Unit",* MATTHIAS MONROY BLOG (Apr. 24, 2019), https://digit.site36.net/2019/04/24/german-police-launches-national-internet-referral-unit/ (reporting based on German Parliament questions procedures).

[65] *See, e.g.*, Marcos Sierra, *Interior se reúne con Google, Facebook y Twitter para la intervención rápida en los delitos de odio* [Interior meets with Google, Facebook, and Twitter about quick intervention in hate crimes], VOZPÓPULI (July 14, 2019), https://www.vozpopuli.com/economia-y-finanzas/interior-google-facebook-twitter-delitos-odio_0_1262574949.html ("'We talked not only that they are hate crimes. We will also alert Twitter about the logos and vignettes that are against the rules of use of the platform. The agreement is interesting because an ad from a company or organization with the 'trusted flagger' mark is analyzed before that from a common user. It has preference. Soon we will meet with Google and Facebook to achieve the same,' explains Carlos Morán, Head of the National Official of the Fight Against Hate Crimes."); *see also* MINISTRY OF INTERIOR – STATE SECRETARIAT FOR SECURITY, ACTION PLAN TO COMBAT HATE CRIMES 16 (2019) ("Appointing the National Office to Combat Hate Speech as 'trusted flagger' for Internet service providers to facilitate the withdrawal of contents including hate speech in coordination with the National Security Forces so there is no interference with on-going judicial investigations. Implementation: second quarter 2019.").

[66] BUNDESAMT FÜR VERFASSUNGSSCHUTZ UND TERRORISMUSBEKÄMPFUNG, https://bvt.bmi.gv.at/601/ (last visited Jan. 12, 2021) ("The BVT has set up a registration office where citizens can notify the Office for the Protection of the Constitution of extremist and radical videos that have a connection to Austria. The BVT will view these videos and initiate appropriate investigations. Furthermore, the videos are reported to the operators, e.g. Google / Youtube. It is up to the operators whether the videos contradict the terms of use and are taken offline by the operator.").

[67] *See, e.g.*, *Referral Action Day with Six EU Member States and Telegram*, EUROPOL (Oct. 5, 2018), https://www.europol.europa.eu/newsroom/news/referral-action-day-six-eu-member-states-and-telegram (mentioning the "the National Referral Units of Belgium, France, Germany, Italy, the Netherlands and the United Kingdom"); *see also* Pielemeier & Sheehy, *supra* note 56.

[68] EUROPOL, *supra* note 67.

[69] *Id.*

[70] *See* EUROPEAN COMMISSION AGAINST RACISM AND INTOLERANCE, ECRI REPORT ON SWITZERLAND (SIXTH MONITORING CYCLE) 18 (2020) ("As for the Internet, ECRI notes the 'flagging mechanisms' introduced by groups such as Facebook and Google which offer the possibility of weeding out fake or offensive content without introducing new laws.54 The authorities informed ECRI that the National Cyber Competence Centre (NC3) of the Federal Office of Police (Fedpol) is seeking cooperation with relevant Internet service providers to improve the identification of authors of hate speech and to have such content removed as quickly as possible. For example, Fedpol's status as 'trusted flagger' allows it to quickly report hate speech content on YouTube to Google, after which the material is taken down rapidly.").

in Germany despite an existing legal apparatus (NetzDG law) that requires companies to independently remove illegal content, highlights the attempts of governments to interpret platforms' own terms of service and takedown content that may not violate national laws.

1. The United Kingdom

The United Kingdom established the world's first IRU in 2010, known as the Counter Terrorism Internet Referral Unit (CTIRU). The unit functions under the Metropolitan Police Service and has been primarily dedicated to identify "terrorist" content on the internet and ask online intermediaries for its takedown. There have been conflicting reports regarding the baseline definition of terrorism and the standard of review that the unit uses to assess and refer content for takedown. Some reports indicate, for example, that the unit refers content that it determines to be in violation of either the UK terrorism legislation or a company's terms of service,[71] while other reports stress that all of the unit's referrals are first assessed against the UK terrorism legislation and are submitted to a company for review only if they violate UK law. The assessment of illegality—if done at all—is internal to CTIRU officials and does not involve judicial review. As the UK Minister of State Security clarifies in response to a parliamentary question: "all referrals are assessed by CTIRU against UK terrorism legislation (Terrorism Act 2000 and 2006). Those that breach this legislation are referred to industry for removal. If industry agrees that it breaches their terms and conditions, they remove it voluntarily."[72]

---

[71] Chang, *supra* note 39, at 129; HL Deb (12 July 2016) (772) col. 8 (UK) ("The police Counter Terrorism Internet Referral Unit (CTIRU) refers content that they assess as contravening UK terrorism legislation or company terms and conditions to Communication Service Providers (CSPs) for removal.").

[72] *Counter-terrorism Question for Home Office*, UK PARLIAMENT (Mar. 14, 2016), https://questions-statements.parliament.uk/written-questions/detail/2016-03-14/30893.

Whether these requests are assessed against the UK terrorism legislation or not, it is remarkable that CTIRU was designed to use an informal path that relies on the companies' discretion, despite existing formal authority to ask for the removal of online content according to the UK terrorism legislation. The Terrorism Act of 2006 authorizes state officials to issue notices to online intermediaries, asking them to takedown content determined to be terrorism-related. Section 3 of the act applies the legal provisions banning the "encouragement of terrorism" or "dissemination of terrorist publication" to content published on the internet.[73] The law allows a "constable" (rather than a judge) to determine that a content is "unlawfully terrorism-related,"[74] and issue a notice that requires that the relevant content "is not available to the public or is modified so as no longer to be so related."[75] A failure to comply with the notice within two working days could account for endorsement of the terrorist content and consequently impose criminal liability.[76]

This authority, however, has never been formally invoked, perhaps due to the wide criticism it faced and its potential inconsistency with the European Convention on Human Rights, as Brian Chang suggested.[77] As social media networks became more central to everyday life, the UK government moved to establish CTIRU which asks for content takedowns based on an informal, voluntary basis. The turn to informality did not prove to be a hurdle, and the unit has been considered enormously successful—so much so that it inspired the establishment of similar units in the EU's Europol as well as across different countries.

---

[73] Terrorism Act 2006, c. 11, § 3(1)(a) (Eng.).
[74] Terrorism Act 2006, c. 11, § 3(3)(a)
[75] Terrorism Act 2006, c. 11, § 3(3)(b)
[76] Terrorism Act 2006, c. 11, § 3(2)
[77] Chang, *supra* note 39, at 127.

Since its inception, CTIRU reportedly succeeded to trigger the removal of "more than 310,000 pieces of extremist material."[78] The UK's Minister of State for Security and Economic Crime has also mentioned that "CTIRU have developed relationships with over 300 online platforms, and are a YouTube trusted flagger."[79] The volume of overall removals, however, did not increase at a linear pace. As an official statement illustrates: "removals at the request of CTIRU have increased from around 60 items a month in 2010, when CTIRU was first established, to over 4000 a month in 2015."[80] A growth in CTIRU removals continued at least until 2018, reporting more than 150,000 removals between 2016 to 2018.[81]

Unlike other IRUs, CTIRU does not publish official annual reports. Existing information regarding its activity is available through occasional reports to media or parliament. While CTIRU has only disclosed statistics regarding the number of *successful*

---

[78] In a 2019 statement, the Head of Counter Terrorism Policing, Neil Basu summarized: "When we first launched out Counter Terrorism Internet Referral Unit (CTIRU) in 2010, it was the first in the world set up to tackle the proliferation of illegal terrorist and violent extremist content online. Since then, it has successfully removed more than 310,000 pieces of extremist material, but that it a drop in the ocean when you consider just how much terrorist propaganda is still available online for those seeking to radicalize themselves and others." *Neil Basu Welcomes Online Safety Measures*, COUNTER TERRORISM POLICING (Apr. 8, 2019), https://www.counterterrorism.police.uk/neil-basu-welcomes-online-safety-measures/.

[79] *Parliamentary Questions, Answer Given by Mr. Ben Wallace, Question Number 70161*, UK PARLIAMENT (Apr. 21, 2017), https://www.parliament.uk/business/publications/written-questions-answers-statements/written-question/Commons/2017-03-30/70161/.

[80] *Counter-terrorism Question for Home Office*, UK PARLIAMENT (Mar. 14, 2016), https://questions-statements.parliament.uk/written-questions/detail/2016-03-14/30893.

[81] *Compare Parliamentary Questions, Answer Given by Mr. Ben Wallace, Question Number 186393*, UK PARLIAMENT (Nov. 8, 2018), https://questions-statements.parliament.uk/written-questions/detail/2018-10-31/18639 ("To date, the CTIRU have secured the removal of over 300,000 pieces of terrorist content, including right wing terrorist content."), *with Parliamentary Questions, Answer Given by Mr. John Hayes, Question Number 30893*, UK PARLIAMENT (Mar. 17, 2016), https://questions-statements.parliament.uk/written-questions/detail/2016-03-14/30893 ("Referrals made to industry by CTIRU have led to over 150,000 pieces of terrorist-related material being removed to date from various online platforms.").

removals, a report by the Open Rights Group suggests that the number of *original referrals* made by CTIRU is considerably higher. According to the report, companies indicated that CTIRU's margin of error lies between 20% to 30%.[82] The report further criticizes CTIRU for its lack of transparency: "The Open Right Group has filed requests for information about key documents held, staff and finances, and available statistics. So far, only one has been successful, to confirm the meaning of a piece of content."[83] This freedom of information response clarifies that counting is administered algorithmically, based on the number of URLs removed, meaning that the removal of an entire website containing multiple pages on WordPress, for example, may account for one removal. The response also implies that CTIRU does not keep documentation of the materials removed: "Please be advised that no documentation exists, but rather a computer generated [sic] model of counting figures."[84]

The lack of transparency regarding CTIRU's work also extends to the lack of oversight mechanisms to review or challenge CTIRU's assessments or referrals apart from the companies' own review processes.[85] The opaque character of CTIRU's work is further complicated by the informal nature of its requests and the purported voluntariness of the removal decisions. A small number of CTIRU requests that found their way to Lumen—an independent research project that collects requests to remove material from the web—demonstrate the informal character of these requests. Some

---

[82] Jim Killock, *Informal Internet Censorship: The UK's Counter Terrorism Internet Referral Unit (CTIRU)*, VOXPOL (July 31, 2019), https://www.voxpol.eu/informal-internet-censorship-the-uks-counter-terrorism-internet-referral-unit-ctiru/.

[83] *Id.*

[84] *Freedom of Information Request*, METROPOLITAN POLICE, https://www.met.police.uk/SysSiteAssets/foi-media/metropolitan-police/disclosure_2018/august_2018/counter-terrorism-command---methodology-used-to-calculate-the-figures-supplied-by-ctriu-to-parliament

[85] Killock, *supra* note 82.

note, for example, that "we would be grateful if using your 'Terms of Use' policy you will block or otherwise restrict access to these pages by for example, removal or suspension."[86] Other requests simply ask "can this please be inspected with a view to removal?"[87] or "can this please be removed?"[88], finally concluding with "thank you for your help, it is really appreciated"[89] or "we are grateful in advance for your cooperation."[90]

2. European Union

Inspired by CTIRU's work, the EU IRU was established in 2015 and is housed within Europol.[91] The referral unit is the only one operating trans-regionally and working to not only refer content to companies but also to advise existing EU Member States' units and encourage others to begin units of their own. The European Union Europol regulation that went into effect in 2017 governs the referral unit's work.[92] In a list of the organization's enumerated responsibilities, one item notes Europol's task of "making of referrals of internet content."[93] The regulation also contains some standards and oversight provisions, but as Brian Chang notes, "the Europol Regulation has been criticized for not doing enough to set

---

[86] *CTIRU Takedown*, LUMEN (Jan. 25, 2016), https://lumendatabase.org/notices/11757313; *CTIRU Takedown*, LUMEN (Jan. 7, 2016), https://lumendatabase.org/notices/11690268.

[87] *CTIRU Takedown*, LUMEN (Sept. 17, 2018), https://lumendatabase.org/notices/17269343.

[88] *CTIRU Takedown*, LUMEN (Oct. 31, 2018), https://lumendatabase.org/notices/17528418.

[89] *CTIRU Takedown*, *supra* note 87.

[90] *CTIRU Takedown*, LUMEN (Jan. 7, 2016), https://lumendatabase.org/notices/11690268.

[91] EUROPOL, EU INTERNET REFERRAL UNIT: YEAR ONE REPORT HIGHLIGHTS 3 (2015); *EU Internet Referral Unit – EU IRU*, EUROPOL, https://www.europol.europa.eu/about-europol/eu-internet-referal-unit-eu-iru.

[92] *See generally* Regulation (EU) 2016/794 of the European Parliament and of the Council of 11 May 2016 on the European Union Agency for Law Enforcement Cooperation (Europol) and replacing and repealing Council Decisions 2009/371/JHA, 2009/934/JHA, 2009/935/JHA, 2009/936/JHA and 2009/968/JHA (L 135/53) [hereinafter "Europol Regulation"].

[93] *Id.* art. 4(m).

the terms for the IRU."[94] Though the directive has data protection standards for the unit, the EU Data Protection Supervisor (the supervisory body) is focused solely on data storage and processing of personal data.[95] The EU legislation discusses a Joint Parliamentary Scrutiny Group to oversee Europol's work, though this oversight feature has generated significant criticism for being politically focused rather than having real authority or capability to inquire into, and report on, Europol and its referral unit.[96]

Europol's IRU follows the European Union's directive on combatting terrorism, and its particular focus is on Islamic State and Al-Qaeda content, though its mandate can extend to any organization on the United Nations Security Council Sanctions List.[97] While the EU IRU's scope for terrorist content is confined to the UN sanctions list, it also refers other non-terrorism material. While CTIRU is only concerned with online "terrorist and extremist" content,[98] the EU IRU is officially authorized to ask for the removal of more than two dozen categories of content, including "immigrant smuggling," "corruption," and "sexual abuse."[99] However, according to its annual reports, the unit seems to primarily focus so far on "terrorist and violent extremist content," as well as "illegal immigrant smuggling networks."[100] The EU IRU highlights that "the unit performed its searches and analysis on material produced in 10 languages, with focus on non-EU languages."[101]

Although the EU IRU has not been operational as long as the UK CTIRU, it has also seen similar explosive growth trends in the volume of content flagged to social media companies for review.

---

[94] Chang, *supra* note 39, at 134.
[95] Europol Regulation, *supra* note 92, at art. 43.
[96] Chang, *supra* note 39, at 199.
[97] Chang, *supra* note 39, at 136.
[98] Chang, *supra* note 39, at 130.
[99] Europol Regulation, *supra* note 92, at annex I.
[100] EU IRU TRANSPARENCY REPORT, *supra* note 46, at 3.
[101] *Id.* at 5.

Comparably, Europol reported assessing a total of 46,392 pieces of "terrorist content" between July 2015 and December 2017, which triggered 44,807 referrals across 170 platforms, with a successful removal rate of 92%.[102] More recent unofficial reports stress that Europol had requested more than 96,160 removals by April 2019.[103]

Each year since 2017, Europol produces its own transparency report.[104] The EU IRU has been officially recognized as a Trusted Flagger by "Google/YouTube and some other OSPs running trusted flagger programmes."[105] Holding this status, as also discussed in Part II A.1, "entails prioritization of content flagged by the EU IRU (Internet Referral Unit)."[106] As a summary describes for their release of the 2019 report, it "give[s] an account of the EU IRU's major activities in 2019," and more specifically, it sheds light on both the prevention activities and the investigative support the EU IRU provided upon request of EU Member States.[107] A brief 11 pages, the document has one section, titled "Referrals" that gives only one aggregate number and is approximately a single page.[108] The rest of the report discusses the legal mandate for the EU IRU, its collaborations with internet companies, and various case studies.

3. France

Following a wave of anti-terrorism and security legislation in France in 2015, the French IRU—officially known as *L'Office*

---

[102] EU IRU TRANSPARENCY REPORT, *supra* note 46, at 5.
[103] Monroy, *supra* note 64.
[104] *EU Internet Referral Unit – EU IRU*, EUROPOL, https://www.europol.europa.eu/about-europol/eu-internet-referal-unit-eu-iru.
[105] Parliamentary Questions, Answer Given by Mr. Avramopoulos on behalf of the commission Question Ref. E-000025/2018, EUROPEAN PARLIAMENT (Mar. 30, 2018), http://www.europarl.europa.eu/doceo/document//E-8-2018-000025-ASW_EN.html.
[106] *Id*.
[107] EUROPOL, EU IRU TRANSPARENCY REPORT (2019), https://www.europol.europa.eu/sites/default/files/documents/eu_iru_transparency_report_2019.pdf.
[108] *Id*. at 6.

Central de Lutte contre la Criminalité liée aux Technologies de l'Information et de la Communication (OCLCTIC)—began assuming expanded powers in not just referring content but also blocking it. First, in February 2015, the French government enacted a law that allowed for the blocking of websites that promote or incite terrorism.[109] Afterwards during the following month, a "subsequent decree provides that the Ministry of the Interior may notify search engines and web directories of content inciting acts of terrorism or justifying them, whereupon the search engines and directories have forty-eight hours in which to delist the content."[110] OCLCTIC does not just report content that may lead to a global removal. In contrast to what ostensibly seems to be completely voluntary decisions by the private intermediary, the French IRU can also employ explicit threats of possible "blocking procedures," namely, geo-block. For example, according to the Lumen database, OCLCTIC notified Automattic, owner of WordPress, that a blog violated the French criminal code. The IRU's complaint clarifies that Automattic had 24 hours to remove the blog and, if it does not comply, then it will enact blocking procedures on the content.[111] Thus the unit's power lies not only in referring content to companies under their terms of service, but also to root out illegal content under French law and take action to block such material within French jurisdiction if platforms refuse to remove it.

---

[109] W. Gregory Voss, *After Google Spain and Charlie Hebdo: The Continuing Evolution of European Union Data Privacy in a Time of Change*, 71 BUS. LAW. 281, 287 (2015-2016); Décret 2015-125 du 5 février 2015 relatif au blocage des sites provoquant à des actes de terrorisme ou en faisant l'apologie et des sites diffusant des images et representations de mineurs à caractère pornographique [Decree 2015-125 of February 5, 2015 on the Blocking of Websites Inciting Acts of Terrorism or Justifying Them and Websites Disseminating Child Pornography], Journal Officiel de la République Française [J.O.] [Official Gazette of France], p. 4168, *available at* http://legifrance.gouv.fr/eli/decret/2015/2/5/2015-125/jo/texte.

[110] Voss, *supra* note 109, at 287.

[111] *French Government Takedown Demand*, LUMEN (Apr. 22, 2016), https://www.lumendatabase.org/notices/12128927.

Despite, or perhaps precisely because of, OCLCTIC's far-reaching authority, the French IRU is the only referral unit of the four examined in this article that maintains a clear internal oversight mechanism. According to this mechanism, the French IRU is supervised by a "designated official" of the *Commission Nationale de l'Informatique et des Libertés* (CNIL). After referring a content for removal by the IRU, the content is simultaneously sent to the CNIL official who can ask the IRU to withdraw its request. If the CNIL and IRU officials dispute regarding a specific removal request, the case can be referred to judicial review of an administrative court, on the discretion of the CNIL official. In February 2019, the administrative court of Cergy-Pontoise ruled for the first time in such a dispute, favoring CNIL position that the content does not qualify for removal.[112] The oversight mechanism still receives criticism that CNIL ultimately "is a purely administrative – and not judicial – institution" and this body "that was founded to implement legislation of importance to the respect of citizens' fundamental rights is now set to become the body that oversees the restriction of citizens' right to freedom of expression."[113] Additionally, in a comment to the Council of Europe about internet blocking practices, one noticed that a CNIL report "highlights that the verification proceedings [where CNIL reviews OCLCTIC requests to remove or block content] were being jeopardized by a lack of resources and by insufficient access to the

[112] Contrôle du Blocage Administratif des Sites: Première Décision Rendue sur Saisine de la Personnalité Qualifiée, CNIL (Feb. 5, 2019), https://www.cnil.fr/en/node/25163.

[113] *France Implements Internet Censorship without Judicial Oversight*, EDRI: BLOG (Mar. 11, 2015), https://edri.org/our-work/france-censorship-without-judicial-oversight/.

relevant information, which in practice makes it difficult to assess whether the requests are well-founded."[114]

The French IRU follows the trend of its peer units with its dramatic yearly increase in removal requests. CNIL, the body that oversees OCLCTIC's work, ultimately does report statistics on the IRU's activity. The CNIL reported in 2018 an increase of 1,270% in content removal requests – from 2,561 requests in between March 2016 and February 2017,[115] up to 35,110 requests in the subsequent year.[116] An additional 18,014 requests were reported in 2019.[117] Most of these requests pertained to content deemed "terrorist,"[118] and a successful removal rate of 78% and 75% was mentioned in 2018 and 2019 respectively. In May 2020, CNIL released its most recent report for activity between February and December 2019. It notes an overall drop in terrorist-related referrals, with a total of 18,177 requests in this window with child sexual abuse material accounting for 68% of the volume.[119]

---

[114] *Arbitrary Internet Blocking Jeopardises Freedom of Expression*, COUNCIL OF EUR.: HUM. RTS. COMMENTS (Sept. 26, 2017), https://www.coe.int/en/web/commissioner/-/arbitrary-internet-blocking-jeopardises-freedom-of-expression.

[115] *Contrôle du Blocage Administratif des Sites: La Personnalité Qualifiée Présente son 2ème Rapport d'activité*, CNIL (May 3, 2017), https://www.vie-publique.fr/sites/default/files/rapport/pdf/184000341.pdf.

[116] *Contrôle du Blocage Administratif des Sites: La Personnalité Qualifiée Présente son 3ème Rapport d'activité,* CNIL (May 30, 2018), https://www.cnil.fr/fr/controle-du-blocage-administratif-des-sites-la-personnalite-qualifiee-presente-son-3eme-rapport (on file with authors).

[117] *Contrôle du Blocage Administratif des Sites: La Personnalité Qualifiée Présente son 4e Rapport d'activité*, CNIL (May 27, 2019) https://www.cnil.fr/fr/controle-du-blocage-administratif-des-sites-la-personnalite-qualifiee-presente-son-4e-rapport.

[118] In 2018, 93% of content was classified as terrorist, while in 2019 only 53% classified terrorist. The report explained that "the decline in the number of OCLCTIC requests for withdrawals of terrorist content is mainly due to the fact that the production of propaganda content by the Daesh terrorist group has fallen sharply." *Id.*

[119] *Contrôle du Blocage Administratif des Sites: La Personnalité Qualifiée Présente son 5ème Rapport d'activité*, CNIL (May 28, 2020), https://www.cnil.fr/fr/controle-du-blocage-administratif-des-sites-la-personnalite-qualifiee-presente-son-5eme-rapport.

4. Israel

The Israeli IRU—officially known as the Cyber Unit at the Office of the State Attorney—is a unit established in 2015 "in view of the need recognized by the State Attorney to coordinate efforts in dealing with crime and terrorism in cyberspace."[120] In addition to advising the prosecution and handling criminal cases in the field of cyberspace, the unit works similarly to other IRUs under what it calls "alternative enforcement." This enforcement includes "acts for removing offensive content, filtering search results, restraining internet users committing forbidden acts," and more.[121] While the Cyber Unit lacks any oversight mechanisms and has admitted to not keeping record of the materials it asks companies to remove,[122] it nonetheless issues annual reports that provide some detail about its work.

Reports published by the Cyber Unit clarify that "alternative enforcement" may be according to one of two "paths": the first is a "formal track," limited to the takedown of entire websites (in contrast to content published on social media for example), where a request is confirmed by judicial authority pursuant to the 2017 Law on Authorities for the Prevention of Committing Crimes Through Use of an Internet Site;[123] the second is a "voluntary track," where the Cyber Unit flags content to internet intermediaries "pointing out a breach of the companies' own terms of service,"[124] and asking for its removal based on this violation. It should be noted, however, that

---

[120] *About the Cyber Unit*, MINISTRY JUST. OFF. ST. ATT'Y (May 20, 2019), https://www.gov.il/en/Departments/General/cyber-about.
[121] Id.
[122] Vicki Alexander, *An Offer They Can't Refuse: Israel's "Informal" Censorship Meets Facebook's Compliance*, SHOMRIM CTR. FOR MEDIA AND DEMOCRACY (Nov. 26, 2020), https://www.hashomrim.org/eng/353.
[123] Law on Authorities for the Prevention of Committing Crimes Through Use of an Internet Site, 5777-2017, SH No. 2650 p. 1040 (Isr.).
[124] ISRAEL MINISTRY OF JUSTICE, CYBER UNIT 2016 ANNUAL REPORT 4 (2017), https://www.justice.gov.il/Units/StateAttorney/Documents/cyber2016.pdf [hereinafter CYBER UNIT 2016 REPORT].

the Cyber Unit purports to refer contents only after it internally determines that the content also violates Israeli law and that there is sufficient "public interest" in acting upon it. These determinations include weighing, *inter alia*, the actual dissemination of the content and its "virality" potential.[125] Since the referred content remains undocumented and there is no external supervisory body in charge of these determinations, it is hard to assess and impossible to contest these decisions.

Similar to apparent trends in other jurisdictions, the volume of content referred by the Israeli Cyber Unit using the "voluntary track" has been rising strikingly. Reports by the Cyber Unit show that whereas in 2016 it submitted 2,241 voluntary takedown requests, in 2017 more than 12,350 requests were submitted—an increase of over 550% within one year. The number of requests continued to grow in subsequent years, with more than 14,280 referrals delivered in 2018 alone, and additional 19,606 in 2019. It is important to note that the amount of contents removed following these requests is in fact much larger, since each request "may contain dozens or hundreds of URLs," as the Israeli State Prosecution report explains, including URLs to full pages or profiles.[126]

While the Cyber Unit says that these referrals may pertain to different types of content, including content deemed by the Unit as harmful to or insulting of public officials, the reports show that the vast majority of referrals—some 98% of overall referrals—have been classified as inciteful or terrorism-related content.[127] In contrast, the vast majority of requests to takedown entire websites

---

[125] HCJ 7846/19 Adalah v. The Cyber Unit ¶ 11 (2021) (Isr.).
[126] CYBER UNIT 2016 REPORT, *supra* note 124, at 5.
[127] Brief of Petitioner at 6, HCJ 7846/19 Adalah v. Office of the State Attorney Cyber Unit (2019), https://www.adalah.org/uploads/uploads/Cyber_Petition_Final_241119.pdf.

based on the "formal track" involved impermissible sexual content, such as child sexual abuse material or sex work content.[128]

The reports also detail the rates of accepted requests and show a growing compliance by the companies: whereas in 2016 the overall compliance rate stood on 76.5%, it grew to 88% in 2017, to 92% in 2018, and stood at 90% in 2019.[129] Some of the reports further break up the referrals according to company, with Facebook—Israel's most popular social media platform—receiving the overwhelming majority of referrals. In 2018, for example, Facebook received 87% of the overall referrals, Twitter received 8% of the referrals, and the rest 5% included platforms such as YouTube, Instagram, Google, and others.[130] The 2019 report details the compliance rate of some companies, citing 82% for Facebook, 93% for Twitter, and 76% for YouTube.[131]

Despite the similarities in the work model between the Israeli Cyber Unit and other IRUs, there remains a crucial difference in the context in which they operate. The employment of the Israeli Cyber Unit cannot be viewed in isolation from Palestinians and invites heightened critical reflection on the takedown of content under the auspices of counter-terrorism.[132] Furthermore, unlike the UK, EU, or France, the Israeli Cyber Unit employs its "voluntary track" without any explicit or implicit legal authorization. This lack of authorization has opened the door for human rights organizations to petition the Israeli High Court of Justice, challenging the informal

---

[128] *See, e.g.*, ISRAEL MINISTRY OF JUSTICE, OFFICE OF THE STATE ATTORNEY 2019 ANNUAL REPORT 58 (2020), https://www.gov.il/BlobFolder/generalpage/files-general/he/DATA%202019.pdf [hereinafter CYBER UNIT 2019 REPORT].
[129] ISRAEL MINISTRY OF JUSTICE, OFFICE OF THE STATE ATTORNEY 2018 ANNUAL REPORT 64 (2019), https://www.gov.il/BlobFolder/generalpage/files-general/he/files_report-2018.pdf [hereinafter CYBER UNIT 2018 REPORT].
[130] Id.
[131] CYBER UNIT 2019 REPORT, *supra* note 128, at 57.
[132] *See infra* Part III.A.2.

activity of the Cyber Unit.[133] The Israeli High Court, however, has recently closed that door in a landmark decision, rejecting to grant relief in the case of *Adalah v. The Cyber Unit*, as discussed further in Part IV.

5. United States

Until this point, the IRUs examined are all officially acknowledged units. However, other countries may either secretly operate such teams or may be thinking about establishing units themselves in a similar vein. In an overview of IRUs across western states one may wonder where the United States fits into this picture. As Brian Chang noted, in 2016 the UK called on the United States to join its efforts and create their own IRU.[134] It is unclear so far whether the US has answered these calls and established a designated office operating officially as an IRU.

In 2015, however, the US House of Representatives passed the Combat Terrorist Use of Social Media Act.[135] The Senate version of the bill, titled the Requiring Reporting of Online Terrorist Activity Act, "would require providers of Internet communications services to report to government authorities when they have 'actual knowledge' of 'apparent' terrorist activity (a requirement that,

---

[133] *Israel State Attorney Claims Censorship of Social Media Content, Following Cyber Unit Requests, Isn't an 'Exercise of Gov't Authority'*, ADALAH –LEGAL CTR. FOR ARAB MINORITY RTS. IN ISR. (Nov. 28, 2019), https://www.adalah.org/en/content/view/9859 [hereinafter ADALAH]. One of the authors of this article is involved in this petition.

[134] Chang, *supra* note 39, at 121 (quoting Theresa May, then Home-Secretary, stating that "I would like to see the United States, Canada, New Zealand and Australia – Britain's Five Eyes Partners – taking the same approach in working with communications service providers to tackle this propaganda."); *see also Home Secretary: International Action Needed to Tackle Terrorism,* GOV.UK (Feb. 16, 2016), https://www.gov.uk/government/speeches/home-secretary-international-action-needed-to-tackle-terrorism.

[135] H.R. 3654, 114th Cong. (2015).

because of its vagueness and breadth, would likely harm user privacy and lead to over-reporting)."[136]

In parallel to the House's passing of the Combat Terrorist Use of Social Media Act of 2016,[137] the White House arranged a closed meeting with tech executives. A leaked document of the meeting's agenda shows that the White House asked to "explore ways to more quickly and comprehensively identify terrorist content online so that online service providers can remove it if it violates their terms of service."[138] It further considered "is there value in creating something similar [to IRUs] here" given that "some governments have undertaken efforts to flag terrorist content online or other terms of service violations for service providers removal."[139] More recently, there have been some indications that the FBI is engaging in similar informal referral behavior, [140] although its scope and link to the 2016 meeting are still unclear.

---

[136] Hugh Handeyside, *Social Media Companies Should Decline the Government's Invitation to Join the National Security State*, ACLU (Jan. 12, 2016, 2:15 PM) https://www.aclu.org/blog/national-security/privacy-and-surveillance/social-media-companies-should-decline-governments.

[137] S. 2517, 114th Congress (2016). The act requires the President to report on the United States efforts to combat "terrorists' and terrorist organizations' use of social media," including "a summary of the Federal Government's efforts to monitor, review, disrupt, and counter the use of social media by terrorists and terrorist organizations." Such report was not found or reported yet.

[138] Danny Yadron & Julia Carrie Wong, *Silicon Valley Appears Open To Helping U.S. Spy Agencies After Terrorism Summit*, GUARDIAN (Jan. 8, 2016), https://www.theguardian.com/technology/2016/jan/08/technology-executives-white-house-isis-terrorism-meeting-silicon-valley-facebook-apple-twitter-microsoft; *White House Briefing Document for Jan. 12 Counterterrorism Summit with Tech Leaders*, INTERCEPT (Jan. 20, 2016) [hereinafter *White House Briefing Document*], https://theintercept.com/document/2016/01/20/white-house-briefing-document-for-jan-12-counterterrorism-summit-with-tech-leaders/.

[139] *White House Briefing Document*, *supra* note 138.

[140] In a short clarifying statement from November 2018, Facebook mentioned that "On November 4, the FBI tipped us off about online activity that they believed was linked to foreign entities. Based on this tip, we quickly identified a set of accounts that appeared to be engaged in coordinated inauthentic behavior, which is banned on Facebook because we want people to be able to trust the connections they make on our services." *See* Facebook, *More Information About Last Week's Takedowns,* FACEBOOK NEWSROOM (Nov. 13, 2018),

Whether the US government has, is planning, or will enact a formal referral unit of its own, there is already evidence that agencies have collaborated with other IRUs. The EU IRU organizes "Referral Action Days" which brings together IRUs across EU member states and sometimes non-EU nations (this is discussed in more detail in Part III below). From a December 2019 press release following one of these action days, Europol writes, "On 27 and 28 November 2019, the European Union Internet Referral Unit (EU IRU) at Europol organized its 17th Joint Referral Action Day with specialized units from EU Member States, non-EU countries and other Europol specialized units."[141] A footnote appended to "non-EU countries" specifies that the FBI was a participant.[142] And even if the US never establishes a unit of its own, it does not mean that it may not attempt to refer content indirectly through an established IRU. It is no surprise that US intelligence agencies collaborate with European and non-European partners, but the lack of context as to the FBI's relationship with the IRU is concerning.

Though this section illustrates general features of IRUs and provides a deeper examination of their operation in five countries, a discussion on the overall risks to speech manifested by the presence and functioning of IRUs must follow. Part III expands on these risks

---

https://about.fb.com/news/2018/11/last-weeks-takedowns/. In this same blog post, the company also noted that "[t]ips from government and law enforcement partners can therefore help our security teams attribute suspicious behavior to certain groups, make connections between actors, or proactively monitor for activity targeting people on Facebook. And while we can remove accounts and Pages and prohibit bad actors from using Facebook, governments have additional tools to deter or punish abuse." *Id.*; *see also* Handeyside, *supra* note 136; Megan Specia, *Facebook Removes Chechen Strongman's Accounts, Raising Policy Questions*, N.Y. TIMES (Dec. 28, 2017), https://www.nytimes.com/2017/12/28/world/europe/chechnya-kadyrov-facebook.html.

[141] *Europol Coordinates Referral Action Day to Combat Manuals and Tutorials on Improvised Explosive Devices Including CBRN*, EUROPOL (Dec. 5, 2019), https://www.europol.europa.eu/newsroom/news/europol-coordinates-referral-action-day-to-combat-manuals-and-tutorials-improvised-explosive-devices-including-cbrn.

[142] *Id.*

of IRUs, which are often hard to capture under the veil of informal governance.

III. INTERNET REFERRAL UNITS AS INFORMAL GOVERNANCE: ASSESSING THE RISKS AND IMPLICATIONS

IRUs employ a system of informal governance: a nonbinding and nontransparent interplay between state actors and private content intermediaries, taking place in the shadow of the law and affecting online speech. The proliferation of IRUs in the past few years invites further scrutiny and assessment of their risks and implication. This chapter embarks on this endeavor and points out main concerns pertaining to their activity.

A. Intrusion on Free Speech and Public Law Norms

IRUs bypass courts and public law frameworks by fostering the takedown of online content under the companies' terms of service. The fact that governments issue these requests to avoid scrutiny is crucial and raises the specter of governmental abuse of powers. The suspicion towards state power is foundational to the initial design of systems that supposedly enjoy from "separation of powers" and "checks and balances." Traditionally, if law enforcement agencies are interested to take down certain content or prosecute the publisher for inciteful content, they would have had to pass through the judiciary. IRUs activity erodes this traditional legal structure, skips judicial scrutiny altogether, and replaces it with private companies' decision-making process.[143] Despite the fact that governments have the option to pursue a legally binding request, they choose to turn to informal engagement to evade constitutional and administrative constraints. Although IRUs working model may reduce governments' costs by "simplifying" the administrative

[143] Private mechanisms of due process are increasingly rising within these platforms. Part IV of this paper considers these mechanisms.

process; it also externalizes the final decision to private companies and avoids crucial oversight, resulting in very limited restraint on IRUs' discretion, if any. In this sense, IRUs manage a new system of governance in which laws are replaced, or rather supplemented, by terms of service and the judiciary by companies' content reviewers.

Resorting to informal governance constitutes a major compromise on transparency and due process, rendering speech removable without the affected person being heard or even knowing that the removal was triggered by a government request in the first place. As long as the content is taken down based on a terms of service violation, rather than formal legal request, companies do not notify users of the identity of the reporter, even if it were a governmental agency. The lack of transparency and due process, coupled with the informal requests that governments submit, not only undermine public law norms but also open the door for governments to censor speech and achieve goals that may otherwise be beyond their legal reach. The absence of transparency, due process, or judicial review, however, are not the only reasons to fear that the takedown decisions may infringe on free speech. There are additional reasons that exacerbate this fear. First, companies' terms of service and national definitions of terrorism are highly susceptible to overbroad interpretations that can result in taking down legitimate content. Second, the unique lawmaking power of the state and the ability of politicians to exert pressure on internet platforms, complicate the very notion of purported voluntariness when it comes to companies' decisions to takedown content. In the following few pages, we explicate on these two factors.

1. Overbroad Interpretation of Terms of Service

Very little is known about the content of IRU requests. The fact that IRU referrals are based on an alleged violation of a

company's terms of service, and most of them allegedly pertain to terrorism-related content, make these takedown requests even more troubling. Consider, for example, the following case. In April 2019, a French governmental unit attracted some attention in the United States when it requested the takedown of more than 550 URLs from the Internet Archive website.[144] The Internet Archive is a non-profit archiving project whose mission "is building a digital library of Internet sites and other cultural artifacts in digital form."[145] These URLs included, to name a few, U.S. government-produced broadcasts and reports, the main page of the Library of Congress hosted by the Internet Archive, scholarly articles, and other user-posted materials, all falsely deemed "terrorist propaganda."[146] Some of these URLs had millions of archived materials on the page.[147]

Just a few days earlier, the French IRU had issued a different request to the Internet Archive, asking them to takedown within 24 hours extensive Quran commentary which it identified as including "provocation of acts of terrorism or apology for such acts."[148] If the content was not removed in time, the Internet Archive could have faced "blocking procedures" in France. These requests were sent via Europol email domains, which pushed the EU IRU to clarify that "it is not involved in the national IRUs' assessment criteria of terrorist

---

[144] For media coverage regarding the French takedown requests, *see, e.g.*, Kalev Leetaru, *The EU's False Terrorist Takedown Requests Remind Us Why Bad Internet Legislation Is So Dangerous*, FORBES (Apr. 12, 2019), https://www.forbes.com/sites/kalevleetaru/2019/04/12/the-eus-false-terrorist-takedown-requests-remind-us-why-bad-internet-legislation-is-so-dangerous/#5453e33baacb; *Internet Archive Denies Hosting 'Terrorist' Content,* BBC NEWS (Apr. 12, 2019), https://www.bbc.com/news/technology-47908220.
[145] *About the Internet Archive*, INTERNET ARCHIVE https://archive.org/about/.
[146] Leetaru, *supra* note 144.
[147] James Vincent, *Archive.org Hit with Hundreds of False Terrorist Content Notices from EU*, VERGE (Apr. 11, 2019), https://www.theverge.com/2019/4/11/18305968/eu-internet-terrorist-content-takedown-mistakes-internet-archive-org.
[148] *Id.*

content."[149] No further information was provided by the French IRU or the EU IRU regarding the false flagging of content, leaving the Internet Archive's staff wondering "are we to simply take what's reported as 'terrorism' at face value and risk the automatic removal of things like THE primary collection page for all books on archive.org?"[150]

While it remains unclear why the French IRU flagged these materials to begin with, it is easy to see that IRU work model opens the door to errors, overbroad interpretations of the companies' terms of service, and the application of national understandings of terrorism onto social media platforms, which can all result in severe ramifications. Scholars have repeatedly warned of the far-reaching censorial effects that may accompany the platforms attempts to regulate online content.[151] This fear is already exacerbated by the "definitional ambiguity"[152] inherent to terms such as "terrorist content," and becomes more pressing when governmental agencies such as IRUs are engaging in the interpretation of terms of service.

The UK's Open Rights Group warns in this context that "we believe the CTIRU had requested removal of extremist material that had been posted in an academic or journalistic context."[153] It further points out to the following example. In a referral that found its way to Lumen, CTIRU requested from WordPress to takedown, under violent extremism concerns, what the Open Rights Group described as "an obviously fake, unpleasant and defamatory blog portraying

---

[149] Chris Butler, *Official EU Agencies Falsely Report More Than 550 Archive.org URLs as Terrorist Content,* INTERNET ARCHIVE BLOGS (Apr. 10, 2019), http://blog.archive.org/2019/04/10/official-eu-agencies-falsely-report-more-than-550-archive-org-urls-as-terrorist-content/.

[150] *Id.*

[151] *See generally* Balkin, *supra* note 4; Citron, *supra* note 6; Klonick, *supra* note 5.

[152] Citron, *supra* note 6, at 1051.

[153] OPEN RIGHTS GROUP, UK INTERNET REGULATION PART I: INTERNET CENSORSHIP IN THE UK TODAY 11 (2018), https://www.openrightsgroup.org/publications/uk-internet-regulation/.

the UKIP party as cartoon figures but also vile racists and homophobes."[154] While WordPress declined this particular request, this example shows that CTIRU employs a broad interpretation to catch unpleasant or borderline content that may nonetheless be legally permissible.

The fear of sweeping censorship, however, is not limited to content deemed "terrorist" or "extremist," and removing certain content may have consequential results beyond intruding on freedom of expression. Brian Chang notes, for example, that the EU IRU "has three full-time staff members dedicated to preventing illegal immigration," which has been criticized for its potential impact on refugees and their safety.[155] Chang further concludes that "this demonstrates the potential for IRUs to be abused, with political leaders initially establishing them to deal with child pornography and to counter violent extremism, but gradually adding new political directives based on the exigencies of the day."[156]

In this context, some may argue that the expanding remit of IRUs and the risk of taking down legitimate content is mitigated by the companies' review process and their voluntary decision to remove content. The fact that a fair percentage of the requests submitted by IRUs do not trigger takedowns could be cited to support the assertion that platforms enjoy independent judgement while reviewing content referred by IRUs. This margin, however, also demonstrates that IRUs are already over broadly interpreting terms of service to catch content that does not violate these terms. While the platforms may still have *some* power to refuse some of the takedown requests, the degree of their purported voluntariness remains deeply in question.

---

[154] *Id.*
[155] Chang, *supra* note 39, at 139.
[156] *Id.* at 140.

2. How Voluntary is Voluntary Removal?

The different IRUs repeatedly highlight that their referrals are non-binding requests but subject to the companies' discretion and voluntary removal. Yet how voluntary is voluntary removal, really? The intermediaries who are asked to make the final decision have conflicting interests and there is a good reason to think they are inclined to comply with state requests at the cost of harming users. Moreover, governments are not only "Repeat Players"[157] that submit thousands of requests through IRUs and have the opportunity to "develop facilitative informal relations with institutional incumbents,"[158] but also possess the power to regulate the legal space in which these platforms function.

The decisions that companies make following informal state requests are therefore made "in the shadows of the law," or in the shadows of the potential laws, while the platforms do not enjoy independence and are clearly subject to asymmetric bargaining powers. The director of the Israeli Cyber Unit, Haim Wismonsky, has, in fact, bluntly acknowledged this:

> Service providers, from their end, fear legislative changes that would expand the state's authority to impose binding orders that would interfere with the ways they moderate content on their platforms. This fear spurs the providers to increase their voluntary cooperation with the states. … The voluntary basis is therefore weakened and it may be viewed as a coercive move from the state, although somewhat concealed.[159]

This does not mean that companies lack the power to refuse IRUs' requests altogether. But this power is certainly subject to a set

---

[157] Galanter, *supra* note 39.
[158] *Id.* at 109.
[159] Haim Wismonsky, *Alternative Enforcement of Illegal Publications Online, in* JUSTICE IN THE LEGAL SYSTEM? CRIMINAL LAW AND CRIMINAL PROCEDURE IN ISRAEL: PROBLEMS AND CHALLENGES 691, 722-3 (Alon Harel ed., 2017) (Hebrew).

of incentives that does not push towards protecting speech. David Kaye, the former UN Special Rapporteur on Freedom of Opinion and Expression, has already noted in this context that "IRUs cannot force platforms to take down content, but they may be vehicles for governments to apply pressure on companies to remove offensive content that doesn't actually violate laws or their terms of service. They take advantage of open-ended platform standards to insist upon takedowns."[160] Experience further demonstrates that forced alignment of corporate and state powers in a space of informality can lead to censorial impact that presumably could not stand constitutional constraints.[161]

This alignment of state and corporate power is most worryingly seen in Israel's efforts to silence and police Palestinian speech. The history of Israel's Cyber Unit and its relationship to social media platforms, most notably Facebook, not only demonstrates this alignment but also its deleterious results.

Following the escalation of violence in October 2015, a period through which Palestinian individuals carried out an uncoordinated series of street stabbings of Israelis (which often resulted in extrajudicial killing of Palestinians), Israeli politicians immediately turned to blame online platforms (alongside Palestinians), and particularly Facebook, for the eruption of violence. Facebook, apart from being the most dominant social media platform in Israel/Palestine, is also registered in Israel and operates offices in Tel Aviv since 2013. In October 2015, then Israeli Prime Minister Benjamin Netanyahu declared that the

---

[160] KAYE, *supra* note 39, at 81.

[161] Take, for example, the case of the successful pressure deployed by politicians to urge Amazon, PayPal, EveryDNS, Apple and others to stop providing services to WikiLeaks, which led to its temporary shutdown. *See* Benkler, *supra* note 35. Additional instances are described by Bambauer, *supra* note 36.

stabbings were the result of "Osama Bin Laden meets Mark Zuckerberg."[162]

Similar attacks on Facebook soon followed. The Interior Minister Gilad Erdan asserted that Facebook "sabotages the work of the Israeli police."[163] Legislation threats immediately followed and a "Facebook bill" that would allow courts to issue orders to takedown content from platforms without an adversarial process was introduced by Erdan and the Israeli Justice Minister at the time, Ayelet Shaked.[164] At the same time, taxation threats against Facebook became more evident and made it to headlines.[165]

In an interview, Shaked explained that following Erdan's announcement, Facebook officials "approached us immediately."[166] Referring to the Israeli IRU that had begun its operation, Shaked said that "today the system is voluntary, if the companies want to, they can takedown the content,"[167] then expressed her dissatisfaction with the current rate of cooperation, which she estimated as 50%. However, Shaked suggested that "[I]t takes time. We met Facebook representatives in Europe two weeks ago and explained all this to them. I am sure that they will eventually understand. We are always in touch with them."[168]

---

[162] Gil Hoffman, *Netanyahu: Palestinian Incitement is 'Osama Bin Laden Meets Mark Zuckerberg,'* JERUSALEM POST (Oct. 19, 2015), https://www.jpost.com/Arab-Israeli-Conflict/Netanyahu-Palestinian-incitement-is-Osama-Bin-Laden-meets-Mark-Zuckerberg-427407.

[163] *Interior Minister Says Facebook a 'Monster', Hindering Security*, YNET (Feb. 7, 2016), https://www.ynetnews.com/articles/0,7340,L-4823259,00.html.

[164] Draft Bill for the Removal of Illegal Content from the Internet, 5777-2016, HH 1104 742 (Isr.), https://fs.knesset.gov.il/20/law/20_ls1_365358.pdf. The bill spurred wide criticism and was eventually blocked by PM Netanyahu on its final stages. *See* Gil Hoffman, *Netanyahu Halts Facebook Bill at Last-Minute*, JERUSALEM POST (July, 18, 2018), https://www.jpost.com/Israel-News/Netanyahu-halts-Facebook-bill-at-last-minute-562802

[165] *See, e.g.*, Steven Scheer, *Israel to Tax Foreign Companies' Online Activities*, REUTERS (Apr. 11, 2016), https://www.reuters.com/article/us-israel-taxation-internet-idUSKCN0X81LJ.

[166] *Ayelet Shaked: Facebook Adayen Lo Mevina et Homrat Hateror Hafalastini*, MAARIV (July 5, 2016), https://www.maariv.co.il/news/politics/Article-548165.

[167] *Id.*

[168] *Id.*

One month later, Jordana Cutler, a former senior advisor to then Prime Minister Netanyahu, was appointed Head of Policy at Facebook Israel offices.[169] Soon after, an additional meeting with Facebook seniors was arranged, and the parties reportedly agreed on "tightening their cooperation."[170] In a conference held immediately after the meeting, Shaked proclaimed that Facebook is increasingly cooperating with Israel and has recently removed 95% of the content it was asked to take down by the IRU, and asserted that "we must keep exerting the pressure, which we will do."[171] The 2016 Israeli Cyber Unit report notes that during 2016, the cooperation with content intermediaries improved significantly and Facebook came to respond to requests "within several hours."[172] The most recent reports of the Israeli Cyber Unit indicate a compliance rate of more than 90%.[173]

These results did not go unfelt by Palestinians who soon complained from a wave of censorship. This wave included the suspension of seven Facebook accounts of journalists from Shehab News Agency and Quds News Network, two main Palestinian news pages that operate on Facebook with over 6 million and 5 million

---

[169] *See, e.g.*, Dorgham Abusalim, *Facebook Hires Longtime Netanyahu Adviser*, MONDOWEISS (June 20, 2016), https://mondoweiss.net/2016/06/facebook-longtime-netanyahu/; Nati Tucker, *Hamenoy Hahadash shel Facebook: Yo'etset Leshe'avar shel Netanyahu* [*Facebook's New Appointment: Netanyahu's Former Advisor*], MARKER (June 13, 2016), https://www.themarker.com/advertising/1.2974413.

[170] In an interview with Cutler, she explained that "ever since I started my job in July, I acted to bring Facebook seniors [to Israel] to listen and develop a deeper understanding of the situation." *See* Tova Tzimoke, S*haked: Facebook Hesira Lebakashateno 95% Tokhni Hasata* [*Shaked: Facebook Removed 95% of Inciting Content Following Our Demand*], YNET (Sept. 12, 2016), https://www.ynet.co.il/articles/0,7340,L-4853699,00.html.

[171] Sharon Pulwer & Elihay Vidal, *Facebook Complying with 95% of Israeli Requests to Remove Inciting Content, Minister Says,* HAARETZ (Sept. 12, 2016), https://www.haaretz.com/israel-news/business/facebook-removes-inciting-content-at-israel-s-request-minister-says-1.5432959.

[172] CYBER UNIT 2016 REPORT, *supra* note 124.

[173] CYBER UNIT 2018 REPORT, *supra* note 129; CYBER UNIT 2019 REPORT, *supra* note 128.

followers respectively.[174] These events led journalists and activists to call for an online protest under the hashtag #FBCensorsPalestine, pushing Facebook to reinstate all seven accounts issuing an apology, confessing that "sometimes we get things wrong."[175]

Additional incidents, however, quickly repeated when Facebook temporarily suspended a page belonging to the Palestinian Information Center (PIC) just one month later, and again in October 2019.[176] Twitter has also suspended without warning or explanation all verified accounts of Quds News Network, suggesting "a clear attack on Palestinian journalism because of Israeli pressure."[177] Similar takedowns have become more common and are occasionally reported by independent Palestinian initiatives and civil society organizations,[178] who have also raised concerns of the Palestinian

---

[174] Sophia Hyatt, *Facebook "Blocks Accounts" of Palestinian Journalists*, ALJAZEERA (Sept. 25, 2016), https://www.aljazeera.com/news/2016/9/25/facebook-blocks-accounts-of-palestinian-journalists. *See also* Peter Baker, *Facebook Struggles to Put Out Online Fires in Israeli-Palestinian Conflict*, N.Y. TIMES (Dec. 7, 2016), https://www.nytimes.com/2016/12/07/world/middleeast/facebook-struggles-to-put-out-online-fires-in-israeli-palestinian-conflict.html.

[175] Hyatt, *supra* note 174.

[176] Marwa Fatafta, *'Incitement' and 'Indecency': How Palestinian Dissent Is Repressed Online*, +972 MAGAZINE (Dec. 4, 2019), https://www.972mag.com/censorship-online-palestinians/.

[177] Ali Abunimah, *Twitter Censors News from Palestine*, ELECTRONIC INTIFADA (Nov. 4, 2019), https://electronicintifada.net/blogs/ali-abunimah/twitter-censors-news-palestine.

[178] Anan AbuShanab, 7AMLEH – THE ARAB CTR. FOR THE ADVANCEMENT OF SOC. MEDIA, *Hashtag Palestine, 2017: Palestinian Digital Activism Report* (2018), https://7amleh.org/wp-content/uploads/2018/04/Palestine-2017-English-final.pdf; 7AMLEH – THE ARAB CTR. FOR THE ADVANCEMENT OF SOC. MEDIA, FACEBOOK AND PALESTINIANS: BIASED OR NEUTRAL CONTENT MODERATION POLICIES? (2018), https://7amleh.org/wp-content/uploads/2018/10/booklet-final2-1.pdf; [This needs an author if there is one, and a date – either of last publication or last access] SADASOCIAL http://sada.social/; Ali Abunimah, *Facebook Labels Palestinian Journalism "Hate Speech,"* THE ELECTRONIC INTIFADA (Mar. 27, 2018), https://electronicintifada.net/blogs/ali-abunimah/facebook-labels-palestinian-journalism-hate-speech. More recently, Palestinians have voiced concern and criticism of a widespread crackdown on online speech regarding forced evictions in the occupied East Jerusalem neighborhood of Sheikh Jarrah. *See* Linah Alsaafin, *Palestinians Criticize Social Media Censorship over Sheikh Jarrah*, ALJAZEERA (May 7, 2021), https://www.aljazeera.com/news/2021/5/7/palestinians-criticise-social-media-censorship-over-sheikh-jarrah.

Authority's recent efforts to adopt similar techniques and foster the takedown of independent journalism and political dissent online.[179]

The coordinated attack on Palestinian online speech has reached new heights during the recent Palestinian uprising against Israeli occupation in May 2021. What had started as criticism against social media censorship of content pertaining to the forced expulsion of Palestinians from the Sheikh Jarrah neighborhood,[180] very quickly escalated to public outcry decrying social media's "digital apartheid" and systemic bias against Palestinians.[181] This included a letter signed by nearly 200 Facebook employees demanding the company investigate the censorship of Palestinian voices on its platforms.[182] Interestingly, Facebook's Oversight Board itself weighed in on the matter in a recent decision, calling for "an independent entity not associated with either side of the Israeli-Palestinian conflict to conduct a thorough examination to determine whether Facebook's content moderation in Arabic and

---

[179] Fatafta, *supra* note 176.

[180] Linah Alsaaafin, *Palestinians Criticise Social Media Censorship over Sheikh Jarrah*, ALJAZEERA (May 7, 2021), https://www.aljazeera.com/news/2021/5/7/palestinians-criticise-social-media-censorship-over-sheikh-jarrah.

[181] *See, e.g.*, Elizabeth Dwoskin and Gerrit De Vynck, *Facebook's AI Treats Palestinian Activists Like It Treats American Black Activists. It Blocks Them.*, WASH. POST (May 28, 2021, 8:09 PM), https://www.washingtonpost.com/technology/2021/05/28/facebook-palestinian-censorship/; Delia Marinescu, *How Facebook, Twitter, and Instagram Have Failed on Palestinian Speech*, SLATE (May 21, 2021, 5:10 PM), https://slate.com/technology/2021/05/dia-kayyali-israel-palestine-facebook-twitter-instagram.html; 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; Kari Paul, *Facebook Under Fire as Human Rights Groups Claim "Censorship" of Pro-Palestine Posts*, THE GUARDIAN (May 26, 2021, 12:00 PM), https://www.theguardian.com/media/2021/may/26/pro-palestine-censorship-facebook-instagram; Omar Zahzah, *Digital Apartheid: Palestinians Being Silenced on Social Media*, ALJAZEERA (May 13, 2021), https://www.aljazeera.com/opinions/2021/5/13/social-media-companies-are-trying-to-silence-palestinian-voices.

[182] Zoe Schiffer, *Facebook Employees Call for Company to Address Concerns of Palestinian Censorship*, THE VERGE (June 1, 2021, 7:27 PM), https://www.theverge.com/2021/6/1/22463952/facebook-employees-petition-palestine-content-moderation-policy.

Hebrew, including its use of automation, have been applied without bias."[183]

The censorship of Palestinian online content had previously spurred suspicion amongst some scholars,[184] linking it to Israel's efforts to battle the BDS movement, potentially through the recruitment of the Cyber Unit methods. As Tamar Megiddo notes, the Israeli Ministry of Strategic Affairs and Information has launched a campaign to fight BDS content online by recruiting and directing users to report content to companies.[185] The website advises users to "share [the content] with us and we will help remove it!"[186]

The case described here not only shows that social media companies are not immune to state power, ultimately manifesting in higher compliance rate with IRU referrals, but also that this phenomenon may result in devastating impact on political speech. The voluntariness of platforms is deeply compromised when state actors exert informal pressure to obtain compliance and cooperation. Israeli Supreme Court Justice Hanan Melcer, who served as the Chair of the Elections Committee during the 2019 Israeli cycles of election, bluntly described a pattern of invoking informal pressure to engender companies' compliance, this time in the context of

---

[183] *Oversight Board Overturns Original Facebook Decision: Case 2021-009-FB-UA*, Oversight Board (Sept. 2021), https://oversightboard.com/news/389395596088473-oversight-board-overturns-original-facebook-decision-case-2021-009-fb-ua/; *see also* Dania Akkad, *Think Facebook's Review Will Fix Alleged Israeli Bias? Not So Fast, Say Palestinians*, Middle East Eye (Sept. 20, 2021), https://www.middleeasteye.net/news/facebook-palestine-israel-oversight-board-bias-fix-not-fast.

[184] *See, e.g.*, Tamar Megiddo, *Online Activism, Digital Domination, and the Rule of Trolls*, 38 COLUM. J. TRANSNAT'L L. 394, 414 (2020) (discussing government efforts to "disrupt the dissemination of undesired information" and mentioning the legal challenge initiated by Adalah).

[185] Megiddo, *supra* note 184.

[186] *Defending Israel Online*, 4IL, https://4il.org.il/ (last visited Sept. 24, 2021).

regulating online speech related to Israeli elections in cooperation with the Cyber Unit:

> I negotiated with Facebook […] They came from the United States, very high ranking officials of Facebook. And then at the beginning, they were very polite, but not responsive. And then I told them that there is another alternative, that if such a case would arise, well, we will go by the legal path […] they asked for several days to think it over. And then they came with a proposed solution, saying that they are willing to apply their procedure of notice and fade out on such matters […] I must tell you that at the beginning, as I told you, they were not so responsive. But after those two weeks, they called it tops. And after that, again, it's not an agreement, but understanding. It went very well.[187]

These informal "understandings" push companies to adapt their moderation and regulation of online speech in a way that aligns with the interests of states and allows the companies to avoid or minimize regulation. The revolving door between governmental offices and private platforms further problematizes this cooperation.[188]

---

[187] Hanan Melcer, Justice, Supreme Court of Israel, *Protecting Elections from Online Manipulation and Cyber Threats: The Experience of Israel's 2019 Election*s, BERKMAN KLEIN CENTER FOR INTERNET AND SOCIETY AT HARVARD UNIVERSITY (Oct. 23, 2019) (transcript available at https://cyber.harvard.edu/sites/default/files/2019-11/2019_10_23_JusticeHananMelcerTranscript.pdf).

[188] In the context of Israel, the appointment of former general director of the Israeli Ministry of Justice, Emi Palmor, to the Facebook Oversight Board, has spurred criticism linking Palmor to the work of the Cyber Unit. *See Palestinian Civil Society Organizations Issue a Statement of Alarm Over the Selection of Emi Palmor, Former General Director of the Israeli Ministry of Justice to Facebook's Oversight Board*, 7AMLEH – THE ARAB CTR. FOR THE ADVANCEMENT OF SOC. MEDIA (May 14, 2020), https://7amleh.org/2020/05/14/palestinian-civil-society-organizations-issue-a-statement-of-alarm-over-the-selection-of-emi-palmor-former-general-director-of-the-israeli-ministry-of-justice-to-facebook-s-oversight-board. *See also* Carlotta Alfonsi, *Taming Tech Giants Requires Fixing the Revolving Door*, KENNEDY SCH. REV. (Feb. 18, 2020), https://ksr.hkspublications.org/2020/02/18/taming-tech-giants-requires-fixing-the-revolving-door/ (describing the movement of individuals between the public sector and the technology industry).

Similar trends of IRU activity in light of larger political pressure employed by high-ranking government officials have also been identified in France and the UK.[189] It is important to note, however, that not all governments possess the same powers to reach, regulate, or influence online intermediaries. While state power may not be the only force at play,[190] differentials in bargaining powers between states also seem to replicate existing global hierarchies.

It is also important to recall that IRU requests may pertain to content published outside their jurisdiction. This extraterritorial activity,[191] initiated by states, makes the use of informal governance even more to the advantage of Western governments and interests. The EU IRU example pertaining to the removal of "illegal immigration services" as well as the Israeli-triggered takedown of

---

[189] Kaye, *supra* note 39, at 79-81. Chang, *supra* note 39.

[190] In some cases, users and civil society organizations may organize and achieve effective bargaining power to pressure the platforms to adopt their demands. A recent controversy that triggered public pressure, for example, concerns Facebook's move to consider policy changes that may redefine the scope of anti-Semitism to include anti-Zionism. Civil society organizations launched a protest campaign and petition under the title "Facebook, we need to talk," which was signed by public figures such as Judith Butler, Cornel West, Noam Chomsky, and more. *See* FACEBOOK, WE NEED TO TALK, https://facebookweneedtotalk.org/ (calling on Facebook not to consider the term "Zionist" as a violation under the company's hate speech policy). While Facebook claimed it did not reach a policy decision on whether to consider the word "Zionist" as a proxy for "Jew," it was later revealed that Facebook had, in fact, already implemented such a policy since 2019. Although this example shows that civil society and users may have some bargaining power, its ability to genuinely change platforms' decision-making processes still remains in doubt. *See* Sam Biddle, Facebook's Secret Rules About the Word "Zionist" Impede Criticism of Israel, THE INTERCEPT (May 14, 2021), https://theintercept.com/2021/05/14/facebook-israel-zionist-moderation/.

[191] *See generally* Bloch-Wehba, *supra* note 39. In 2019, the European Union's Court of Justice ruled that Member States have the authority to order platforms to remove content globally. *See* Case C-18/18, *Glawischnig-Piesczek v. Facebook Ireland Limited*, ECLI:EU:C:2019:821, ¶ 53 (Oct. 3, 2018) ("article 15(1) [of directive 2000/31], must be interpreted as meaning that it does not preclude a court of a Member State from … ordering a host provider to remove information … worldwide within the framework of the relevant international law."); *see also* Adam Satariano, *Facebook Can Be Forced to Delete Content Worldwide, E.U.'s Top Court Says*, N.Y. TIMES (Oct. 3, 2019), https://www.nytimes.com/2019/10/03/technology/facebook-europe.html.

Palestinian journalistic content both show how IRU activities are likely to affect already-marginalized groups.[192]

These differences in the bargaining power of states in relation to companies can be traced to several important factors. While some governments can effectively impose and enforce regulation on big tech companies, and even tax them if they operate offices within their jurisdiction, other governments find obstacles in even contacting these platforms.[193] Additionally, as Klonick points out, the main social media companies were developed in the US and therefore carry the American legal compass within them. Countries from the Global North with higher political and economic leverage are more likely to be able to influence these companies' policies, since companies' decisions are likely to be affected by the consumerist power of the market base within a specific country, which in turn may replicate existing socio-economic inequalities.

B. Company Adoption of State Interpretation of Terms of Service

IRUs and the law enforcement agencies in which they operate have been directly involved in defining impermissible speech on private platforms, a task that normally rests with the companies. Although companies can, and occasionally do, reject some IRU interpretations, each referral that is submitted is an alternative reading of the companies' often vague community guidelines around terrorist or extremist content. This "definitional ambiguity"[194] can lead to over-flagging content that IRUs feel violate a company's terms of service. There are already significant differences in how each country defines and labels terrorism in its own legislative apparatus; the companies themselves also think

[192] Chang, *supra* note 39, at 123.
[193] Hamilton, *supra* note 45.
[194] Citron, *supra* note 6, at 1051.

about their respective terrorism policies differently from one another, offering broad and poorly scoped terms in their guidelines.[195] Even if IRUs and companies do not differ significantly in how they think about "terrorist" content online, even shades of alternative interpretations can be harmful. As one critique of the CTIRU put it, "true relationship between CTIRU content removals and matters of national security and crime preventions is likely to be subtle, rather than direct and instrumental."[196] Tens of thousands of IRUs referrals constitute a continuous, open invitation for companies to adopt state interpretations of their own terms of service.

These trends of public-private partnership illustrate the atmosphere in which informal governance takes place, and the risks that IRUs may entail to free speech, which become more pressing given the lack of sufficient transparency and due process. By threatening to invoke regulatory powers, governments are tightening their cooperation with intermediaries and paving the way for IRUs to signal specific contents for takedown and simultaneously engage with the companies in an interpretative task on the boundaries of online speech. This interplay is not only resulting in the direct censorship of some content, but is also enabling governments to affect, from the bottom-up, the interpretation of the usually vague and overbroad "community standards" and terms of service. In other words, the risks of IRUs are not confined to the direct violation of free speech and due process requirements as a result of their activity, but also expand to the ability of such domestic units to engage with

---

[195] *See generally Violent Organizations Policy*, TWITTER HELP CTR. (Oct. 2020), https://help.twitter.com/en/rules-and-policies/violent-groups; *Dangerous Individuals and Organizations Policy*, FACEBOOK COMMUNITY STANDARDS, https://www.facebook.com/communitystandards/dangerous_individuals_organizations; *Violent Criminal Organizations*, YOUTUBE HELP CTR., https://support.google.com/youtube/answer/9229472?hl=en.

[196] Killock, *supra* note 82.

terms of service and shape the limits of what is considered legitimate speech in the global virtual arena, across social media platforms.

This process of offering companies alternative interpretations of their own terms of service appears to be working. The UK's Detective Chief Superintendent Southworth remarked in an interview:

> At its height around 2016, the CTIRU was identifying and removing around 10,000 pieces of content every month, but this has reduced considerably in the past year. It's not because the material is no longer there, but that others are now stepping up and taking more responsibility—particularly major internet service providers. It's a reflection of the excellent work and dedication of the officers in the CTIRU and their efforts to both raise awareness of this issue across the industry, and export the CTIRU model to other countries around the world that we're seeing the tide turn against terrorists online.[197]

These remarks of platforms "taking more responsibility" and the spread of the "CTIRU model" are indicative of this larger trend. IRUs are not only engaging in an exercise that contests specific posts, tweets, or videos; they are also involved in a more subtle effort of getting others to see the issue the way they do. And it seems to be working. More recently, the UK's CTIRU seems to have been shifting direction in its work: "Technology companies and providers are also now becoming more effective at removing the content themselves, which has allowed for a shift in focus for the unit, from removals to investigations."[198] This shift, involving the companies' "successful" interpretation and enforcement of their terms of service, suggests that IRUs' are not only flagging particular pieces

---

[197] *Together, We're Tackling Online Terrorism*, UK POLICE (Dec. 19, 2018), https://www.counterterrorism.police.uk/together-were-tackling-online-terrorism/

[198] *Id.*; *see also Neil Basu Welcomes Online Safety Measures*, UK POLICE (Apr. 8, 2019), https://www.counterterrorism.police.uk/neil-basu-welcomes-online-safety-measures/.

of content to be removed on voluntary basis, but are also engaging in "training" companies on how to interpret their own terms of service. Even if IRUs halted all referrals abruptly, it's unclear the extent to which companies may have adopted IRU thinking to become "more effective," so much so that CTIRU's focus has changed.

In the same statement, the CTIRU also mentioned "[t]his has been achieved through initiatives, such as the Global Internet Forum to Counter Terrorism" a governance structure that warrants additional explication. IRUs have increasingly offered their own interpretations of companies terms of service through new public-private collaborations with the private sector. The EU IRU's work on promoting hash-sharing initiatives—most notably the Global Internet Forum to Combat Terrorism (GIFCT) hash database—and its extensive cooperation with tech companies through "Referral Action Days" shows how IRUs are not only engaging in terms of service interpretative exercises but are also intervening directly in company efforts to define terrorist content.

The EU IRU has long been enthusiastic about the potential of the hash-sharing database conducted by the Global Internet Forum to Combat Terrorism. Hashes, or "digital fingerprints," belong to unique posts, videos, or images on a site that can then be used to search for identical copies on other platforms. YouTube, Facebook, Microsoft, and Twitter created a database of hashes as part of their creation of the Global Internet Forum to Combat Terrorism (GIFCT). This repository is composed of content removed by social media companies for violating a company's specific policies on terrorism.[199] Any GIFCT-participating company

[199] Kent Walker, *To Stop Terror Content Online, Tech Companies Need to Work Together*, GOOGLE: KEYWORD (Dec. 20, 2018), https://www.blog.google/outreach-initiatives/public-policy/stop-terror-content-online-tech-companies-need-work-together/

can then check its corpus of content against the database and contribute to it as well. The ultimate goal of this database is to prevent the emergence of content removed by one platform on other social media sites.

Appetite for hash-sharing databases has only grown in the European context. Europol's unit participates in the EU Internet Forum which brings together government agencies and technology companies to address issues facing the industry.[200] After mounting criticism about the proliferation of terrorist content on YouTube and other social media platforms, YouTube, Facebook, Microsoft, and Twitter announced their hash-sharing initiative at a meeting of the forum.[201] The EU IRU has commented in its Annual Activity Report for 2018 that it "provided relevant contents to feed the database of hashes."[202] The GIFCT's Transparency Report maintains that "[t]here have been no formal requests from Law Enforcement or Governments to gain access to the hash sharing consortium database."[203] Though a referral unit may not be able to *examine* the contents of the database, there remains the opportunity for IRUs to pressure certain types of content to be added to a centralized database.[204] As the Center for Democracy and Technology (CDT) notes, the European Commission itself in its proposed regulation on online terrorist content "calls for 'working arrangements between all relevant parties, including where appropriate Europol' to ensure 'a

---

[200] Chang, *supra* note 39.

[201] *EU Internet Forum: A Major Step Forward in Curbing Terrorist Content on the Internet*, EUR. COMMISSION (Dec. 8, 2016), https://ec.europa.eu/commission/presscorner/detail/en/IP_16_4328.

[202] Europol, 2018 Consolidated Annual Activity 44 (2019).

[203] GIFCT, TRANSPARENCY REPORT 3 (2019), https://gifct.org/wp-content/uploads/2020/10/GIFCT-Transparency-Report-July-2019-Final.pdf.

[204] Civil society groups have criticized centralized content databases for their high risk of being abused. *See, e.g.*, Emma Llansó, *Takedown Collaboration by Private Companies Creates Troubling Precedent*, CTR. FOR TECH. AND DEMOCRACY (Dec. 6, 2016), https://cdt.org/insights/takedown-collaboration-by-private-companies-creates-troubling-precedent/.

consistent and effective approach' to content removal through the database."[205] Emboldened by procedural departures from traditional legal removal requests, Europol and other law enforcement actors may help influence the composition of the database. In a 2020 letter, a group of more than ten human rights NGOs, including AcessNow, Amnesty International, the Center for Democracy & Technology, Human Rights Watch, the Syrian Archive, and more, addressed the GFICT, warning from a series of risks tied to its work, including harming free speech and removing valuable documentation of human rights abuses. The letter additionally notes that "GIFCT is also engaging with law enforcement and experts in challenging violent extremism and counter-terrorism without transparency or any real assessment of the potential human rights harms this could cause."[206]

Though the EU has discussed the GIFCT database publicly, some of the conversations about the database and its contents may have happened privately. The EU IRU has not only tried to define impermissible speech by focusing on the hash-sharing database, it also meets regularly with companies. Since at least 2016, the unit has organized "Referral Action Days" which bring together company representatives with members of the EU IRU and member state IRUs. In one press release, Europol noted that in 2019, its 16th event, "was joined by a total of 9 online service providers, including Telegram, Google, Files.fm, Twitter and Instagram."[207] As part of

---

[205] Emma Llansó, *Who Needs Courts? A Deeper Look At the European Commission's Plans to Speed Up Content Takedowns*, CTR. FOR TECH. AND DEMOCRACY (Mar. 1, 2018), https://cdt.org/insights/who-needs-courts-a-deeper-look-at-the-european-commissions-plans-to-speed-up-content-takedowns/.

[206] Emma Llansó, *Human Rights NGOs in Coalition Letter to GIFCT*, CTR. FOR TECH. AND DEMOCRACY (July 30, 2020), https://cdt.org/insights/human-rights-ngos-in-coalition-letter-to-gifct/.

[207] *Referral Action Day Against Islamic State Online Terrorist Propaganda*, EUROPOL (Nov. 22, 2019), https://www.europol.europa.eu/newsroom/news/referral-action-day-against-islamic-state-online-terrorist-propaganda.

the event, "representatives from Google and YouTube came to Europol's headquarters for an expert exchange, continuing the companies' ongoing cooperation in tackling terrorism online since the creation of the EU IRU."[208] Facebook has also been highlighted as a participating industry member in other events.[209] Because records or summaries of these private meetings are not shared, the full context of what is and is not discussed remains to be fully understood. But given these activities, it becomes clear that IRUs are not only engaged in particular referrals, but also in interpreting what is to be considered impermissible content and training companies to interpret their own terms of service accordingly.

C. Impeding Human Rights Documentation Efforts

As discussed above, there is significant ambiguity and lack of consistency in terrorism definitions across the platforms' community guidelines. Even across countries, "the absence of a universally agreed definition of terrorism presents an ongoing obstacle to any internationally agreed approach to the appropriate regulation of terrorism-related activity and content over the Internet."[210] The range of interpretative possibilities can fuel IRUs' specific conclusions whether a particular piece of content is or is not terrorist material. This last risk threatens to impede human rights documentation efforts, especially since much of this "terrorist" content could actually serve critical evidentiary value of ongoing

---

[208] *Id.*

[209] *EU Law Enforcement Joins Together with Facebook Against Terrorist Propaganda*, EUROPOL (Jan. 12, 2018), https://www.europol.europa.eu/newsroom/news/eu-law-enforcement-joins-together-facebook-against-online-terrorist-propaganda.

[210] U.N. OFFICE ON DRUGS AND CRIME, THE USE OF THE INTERNET FOR TERRORIST PURPOSES 30 (2012), https://www.unodc.org/documents/frontpage/Use_of_Internet_for_Terrorist_Purposes.pdf.

abuse in a particular country or region.[211] Since much of the IRUs' referrals focus on terrorism content, this material may also include content either uploaded by human rights actors but mistook to be "terrorist" or provide important evidence into abuses of power, crimes against humanity, and other human rights violations.

Even if one sets aside the spectrum of definitions and labels that countries and companies alike employ to define terrorism, it is just as important to consider the defensive posturing and language these institutions use when talking about terrorist content. This tough-on-terrorism rhetoric belies the blunt force approach these actors take, casting as wide a net as possible which inevitably harms human rights actors. As a report from the Electronic Frontier Foundation, the Syrian Archive, and WITNESS puts it so succinctly, "[b]lunt content moderation systems at scale inevitably make mistakes, and marginalized users are the ones who pay for those mistakes."[212]

Social media content is increasingly playing a larger role in international criminal law and other investigatory efforts into human rights abuses. At the same time, this content also serves as a vital link for displaced individuals and for diasporic communities to preserve these histories and voice trauma.[213] For example, the International Criminal Court's arrest warrant for Mahmoud Mustafa Busayf Al-Werfalli in 2017 was the ICC's first warrant issued based primarily on social media evidence.[214] However, in this "mad rush

---

[211] Amre Metwally, *The Context Problem Social Networks Don't Like to Talk About*, SLATE (Dec. 17, 2020), https://slate.com/technology/2020/12/facebook-oversight-board-context-human-rights.html.

[212] Abdul Rahman Al Jaloud et al., ELECTRONIC FRONTIER FOUNDATION, *Syrian Archive & WITNESS, Caught in the Net: The Impact of "Extremist" Speech Regulations on Human Rights Content* 6 (2019), https://www.eff.org/files/2019/05/30/caught_in_the_net_whitepaper_2019.pdf.

[213] *Id.* at 7.

[214] Emma Irving, *And So It Begins… Social Media Evidence in an ICC Arrest Warrant*, OPINIOJURIS (Aug. 17, 2017), http://opiniojuris.org/2017/08/17/and-so-it-begins-social-media-evidence-in-an-icc-arrest-warrant/

to 'eliminate' poorly defined 'terrorist and violent extremist content'"[215] there is already ample evidence of erasing human rights material.[216] A recent Human Rights Watch report in September 2020 on the deletion of human rights content even brings IRUs into its discussion. The report says:

> Compounding concerns over the disappearance of potentially valuable online evidence, a growing number of governments, as well as Europol, have created law enforcement teams known as internet referral units (IRUs) that flag content for social media companies to remove, with scant opportunity to appeal or transparency over the criteria they use and how much of the removed material they archive, if any.[217]

An Alphabet employee even referred to the Syrian conflict as "[t]he Syrian civil war is in many ways the first YouTube conflict in the same way that Vietnam was the first television conflict."[218] As more turn to social media in current conflicts or incidents of human rights violations (and surely ones to come), the risk of taking down this material under the umbrella of extremism, violent, or terrorist content, triggered by an IRU, remains high.

---

[215] Dia Kayyali, *WITNESS Joins 14 Organizations to Urge GIFCT to Respect Human Rights*, WITNESS BLOG (Jul. 30, 2020), https://blog.witness.org/2020/07/witness-joins-14-organizations-to-urge-gifct-to-respect-human-rights/.

[216] Dia Kayyali & Raja Althaibani, *Vital Human Rights Evidence in Syria is Disappearing from YouTube*, WITNESS BLOG (Aug. 2017), https://blog.witness.org/2017/08/vital-human-rights-evidence-syria-disappearing-youtube/; Kate O'Flaherty, *YouTube Keeps Deleting Evidence of Syrian Chemical Weapon Attacks*, WIRED (June 26, 2018), https://www.wired.co.uk/article/chemical-weapons-in-syria-youtube-algorithm-delete-video; *Social-media Platforms Are Destroying Evidence of War Crimes*, ECONOMIST (Sept. 21, 2020), https://www.economist.com/international/2020/09/21/social-media-platforms-are-destroying-evidence-of-war-crimes.

[217] HUMAN RIGHTS WATCH, "VIDEO UNAVAILABLE": SOCIAL MEDIA PLATFORMS REMOVE EVIDENCE OF WAR CRIMES 11 (2020), https://www.hrw.org/report/2020/09/10/video-unavailable/social-media-platforms-remove-evidence-war-crimes.

[218] Armin Rosen, *Erasing History: YouTube's Deletion of Syria War Videos Concerns Human Rights Groups*, FAST COMPANY (Jul. 3, 2018), https://www.fastcompany.com/40540411/erasing-history-youtubes-deletion-of-syria-war-videos-concerns-human-rights-groups.

IV. INFORMAL GOVERNANCE CONTESTED: TOWARDS LEGAL CONSTRAINTS ON STATE REFERRALS

With little to no legal constraints over their activity, IRUs are fostering informal governance and generating significant concerns as they proliferate. However, the turn of governments to informal work models structured around the private companies' terms of service has also started to acquire some legitimacy. In a decision delivered recently, the Israeli Supreme Court rejected a challenge to the Cyber Unit's referral activity and ruled that the practice is consistent with constitutional and administrative standards. Looming over the legal dispute stood the issue of censoring Palestinian speech. In addition to this impact, the case remains the first and only decision worldwide to grant a green light for the operation of an IRU.[219] It therefore provides an opportunity to analyze the compatibility of informal structures of governance with constitutional norms and scrutinize the court's reasoning as more legal cases may arise.

The petition, submitted by Adalah—The Legal Center for Arab Minority Rights in Israel and the Association for Civil Rights in Israel, argued that the referral activity of the Cyber Unit violates constitutional norms, including freedom of expression and due process, without any authorization by law. While the lack of statutory authority in Israeli law provided the initial legal hook to challenge the Cyber Unit, the petitioners also argued that the threat of imposing regulation and taxation on the platforms, many of

[219] The concurrence in *Davison v. Randall* raises this question yet leaves it unresolved: "while a government official, who under color of law has opened a public forum on a social media platform like Facebook, could not ban a user's comment containing hate speech, that official could report the hate speech to Facebook. And Facebook personnel could ban the user's comment, arguably circumventing First Amendment protections. Admittedly, this question is not directly presented in the present case, given that the public official, not a Facebook employee, acted to restrict speech. *Supra* note 13, at 45-6.

whom are incorporated in Israel (e.g. Facebook and Google), substantially weakens the claim for voluntariness. In response, the Cyber Unit claimed that the referrals do not constitute state action since they are ultimately subject to the private platforms' discretion. Put differently, the government argued that referrals and removals should not be distinguished; as long as the removal is subject to the private platforms' discretion, then the referral does not constitute an independent state action that requires explicit authorization or compliance with public law constraints.

While the court ruled that the referrals are state action since they may "influence the discretion of the content intermediaries,"[220] it nonetheless found that the activity is "crucial to the national security and social order" and is sufficiently grounded in the government's residual authority and general policing powers. [221] This conclusion comes after the court found that the petitioners failed to link the Cyber Unit's referrals to concrete violation of rights. The court went further and assumed that a bulk of the content referred is generated by bots, stating that "robots do not have human rights."[222] As for content posted by human individuals, the court noted that "as long as a violation [of rights] exists, it is carried out by the operators of the online platforms, and not by a state actor."[223] Interestingly, the decision mentions in this context the Facebook Oversight Board as a possible forum that may provide remedies for the violation of rights, granting a substantial stamp of legitimacy to private platforms' semi-judicial bodies and processes. Finally, the court suggested, without requiring it, that the Cyber Unit should consider establishing an oversight mechanism over its activity and improve its transparency reports to include select examples of the

[220] Adalah v. The Cyber Unit, *supra* note 41, ¶ 53.
[221] *Id.* ¶ 72.
[222] *Id.* ¶ 31.
[223] *Id.* ¶ 67.

contents referred. In short, the court acknowledged IRU referrals as state action without imposing constitutional or administrative constraints that a state action is traditionally subjected to.

An additional important aspect of the court's analysis pertains to the understanding of the state's informal engagement with the private platforms' terms of service. While the decision acknowledges that the state's regulatory power may influence the platforms' decisions, it simultaneously goes on to endorse the state's engagement with the platforms' self-regulation apparatus, dubbing it "reverse regulation." According to the court, this "new" type of regulation allows the state to "fulfill its duty to prevent criminal offenses in a fast and efficient way" by "subjecting itself (ostensibly) to the decisions of market players."[224] Reverse regulation allows the state to "influence and facilitate the action of market players" through engaging with their existing self-regulation mechanism instead of turning to direct regulation. In other words, through the introduction of "reverse regulation," the court in fact endorsed informal governance as a desired and efficient solution.

The court's logic and its endorsement of "fast" and "efficient" regulatory models, however, stops short of asking what actions are actually being taken under the pretext of informal law enforcement, what protections are eroded along the way, and at whose expense. The court's conclusion, in this context, that referrals have not been proven to result in a specific violation of rights further ignores the fact that proving such violation is impossible by design. Since IRUs rely on the platforms' terms of service and since the content of these interactions is not subject to any meaningful transparency or oversight (neither by the governments nor by the private platforms), it is practically impossible to know which removals are triggered by state referrals.

---

[224] *Id.* ¶ 55.

Furthermore, the court's analysis advances a faulty understanding of the hybrid nature of informal governance: on one hand it acknowledges the state's leverage to influence market actors as they interpret their own terms of service, on the other hand it reinforces the division between the public and private spheres while leaving the administrative body and the private platforms as the only guards of their own conduct in their own separate spheres. But even without altering the legal division between the public and private spheres, the court's analysis does not consistently treat referrals as distinct from removals and blur the understanding of these actions as a product of the public and private divide, respectively. If it did, the conclusion that state referrals are state action should have entailed subjecting them to public law constraints such as formally enforceable due process, sufficient transparency and public oversight mechanism.

More generally, it remains important to stop and ask what legitimacy informal governance enjoys? When state referrals and platforms' removals are enmeshed together under the veil of informality, the legitimacy of these takedowns becomes harder to justify based on theories of will or reason. The problem of legitimacy becomes even more pressing when state actors are involved in the takedown of content published beyond their jurisdiction. The overall legitimacy of IRUs should also be considered against the lack of meaningful oversight over their activities. In this context, private law oversight mechanisms that are now emerging, with Facebook's launch of its Oversight Board, cannot provide, in their current form, a sufficient answer for state-initiated censorship. Besides the fact that the legitimacy of these institutions remains highly questioned, they are neither designed to

provide meaningful due process in individual cases[225] nor can effectively restrict states from exploiting the private self-regulation structure.

Other existing mechanisms of internal oversight based on the French IRU model have also proved to be insufficient to constitute major checks. Although the French IRU comes closest to enabling a review of the referral by setting internal oversight mechanism, this mechanism is merely available *after* a referral has already been made and is only subject to the discretion of the administrative branch itself. The low number of referrals reviewed sheds further doubt on the efficacy of this mechanism. Ultimately, citizens/users as well as the broader public are left without effective opportunity for active participation in review.

It should be clarified: while this article does not view informal governance in absolute normative terms, it does approach the issue with a healthy dose of skepticism. Informal governance is a framework that highlights the liminal public/private space within which interactions between state actors and private platforms occur, brings into view their ramifications, and complicates our understanding of the legal infrastructure governing the content moderation enterprise. The risks discussed in the previous section inform our understanding of the phenomena in the context of IRUs. It becomes clear that employing informal governance through IRUs has a substantial cost that cannot simply be dismissed along the lines of "trust us, we're doing the right thing" or "this is more efficient." Furthermore, there is a basis to believe that these units are driven by a sense of false-necessity, or at the least that this necessity of

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

"helping" corporations "find" the content that should be taken down, can be sufficiently met by fostering non-state watchdogs.[226]

The persistence of IRUs, however, invites reflection on possible ways to mitigate the risks they pose. While collapsing the public/private distinction and applying public law norms on private platforms may not be desirable,[227] maintaining this distinction should at least entail subjecting state referrals to effective public law constraints including explicit authorization by law, meaningful transparency, and due process.[228] Allowing states to engage informally with existing structures of private governance under the pretext of efficiency does not solve the threat of a likely alignment between state and corporate power at the expense of the weaker part of the triangle: citizens/users. As long as informal governance is built on this dichotomy between the public and the private, states should not be able to enjoy a lower standard of review by utilizing the contractual agreements between users and platforms and "hide" behind private actors.

---

[226] The option of a state watchdog taking this responsibility raises many of the same concerns that are highlighted with IRUs. There are, of course, state watchdogs that oversee print and television media. Recently, the UK government decided to appoint its communications regulatory body, the Office of Communications (Ofcom), to now also monitor social media platforms and "primarily make sure social networks enforce their own terms and conditions." Alex Hern, *What Powers Will Ofcom Have to Regulate the Internet?*, GUARDIAN (Feb. 12, 2020), https://www.theguardian.com/media/2020/feb/12/what-powers-ofcom-have-regulate-internet-uk. Such cooperation with independent organizations, however, seems to avoid the scenario of compelling platforms to remove content that a government deems objectionable, or else face the threat of regulation.

[227] Such application is likely to spur a whole new set of contradictions between national laws' differing standards of free speech, and require online intermediaries to shift from universalized to localized standards—a Balkanization that could threaten how we all use the internet to communicate. *See* A. Michael Spence, *Preventing the Balkanization of the Internet*, COUNCIL ON FOREIGN REL.: RENEWING AMERICA (Mar. 28, 2018), https://www.cfr.org/blog/preventing-balkanization-internet; *see also supra* text accompanying note 20.

[228] Whereas compliance with constitutional and administrative law constraints is traditionally enforced through domestic judicial systems, Brian Chang has also suggested that the European Court of Human Rights (ECtHR) may prove to be helpful in challenging the conduct of IRUs. Chang, *supra* note 39, at 178-79.

On behalf of the users that rely on these social media platforms, companies should provide proper notifications to their users when content has been actioned as a result of an IRU referral. While content that is removed or restricted following legal takedown requests is marked as such,[229] content removed based on an IRU request is not distinguishable from ToS removals more broadly. If a post, Tweet, or video is removed as the result of state interference, users should be able to know. At the very least, platforms should include content removed based on IRU referrals far more explicitly in their transparency reporting.

What becomes clear overall, however, is that states should not be able to avoid public law constraints by turning to informality. In this context, imposing limitations by requiring IRUs to satisfy a formal procedure and pass through the judicial system, does not only promote formal governance in the sense that it is less opaque, but also obliges states to indirectly pay a cost for initiating censorship through internalizing the procedural costs, rather than externalizing them to private corporations. The panic of governments over potentially unlawful online speech is unprecedented in comparison with offline speech, and on their way to achieve a "sterile" online environment, governments are intruding on virtual civic space and function much more than merely law enforcement authorities under blurred informality.

[229] For example, Twitter marks Tweets as "country withheld content" if the company geo-blocks this material on the basis of its own review of a legal removal request. *About Country Withheld Content*, TWITTER RULES AND POLICIES, https://help.twitter.com/en/rules-and-policies/tweet-withheld-by-country (last visited May 7, 2021).

CONCLUSION

As IRUs continue to grow—in terms of the number of countries that adopt this model *and* the overall number of content flagged—it is abundantly clear that governments see the advantages this model offers to bypass formal legal paths and constraints. Informal collaborations and partnerships between governments and private intermediaries existed before IRUs emerged, though this form of oversight and control was originally *ad hoc* bursts and sputters. Now, however, the proliferation of IRUs solidifies the rise of informal governance in a more systemized manner, establishing relationships with platforms to allow the flagging and subsequent removal of tens of thousands of videos, posts, and tweets each year. Furthermore, informal governance is advancing a new mode of cooperation that relies on interpreting and enforcing the private platforms' own terms of service in an ostensibly voluntary fashion.

The censorial impact ultimately extends beyond discrete pieces of content. Over time, states are able to shape, color, and contour the margins and "grey areas" of these platforms' content policies, signaling to companies what exactly a desirable interpretation of their terms of service should be. In other words, Facebook, YouTube, and Twitter have a new enforcement branch that stretches outside of their content moderation workforce, with IRUs scouring platforms for material that should be enforced against terms of service. Besides going against basic principles of separation of powers and engaging in informal activity that transcends the rule of law, IRU activity threatens marginalized groups and voices. In fact, the brunt of this enforcement continues to fall on them. The rise of this system of informal governance cannot be ignored and its legitimacy must be contested—publicly and legally. Otherwise we allow a handful of states to continue to informally define the boundaries of online speech worldwide, one flag at a time.