ALLOCATING RESPONSIBILITY IN CONTENT MODERATION: A FUNCTIONAL FRAMEWORK

Deirdre K. Mulligan† *& Kenneth A. Bamberger*††

ABSTRACT

This Article develops a framework for both assessing and designing content moderation systems consistent with public values. It argues that moderation should not be understood as a single function, but as a set of subfunctions common to all content governance regimes. By identifying the particular values implicated by each of these subfunctions, it explores the appropriate ways the constituent tasks might best be allocated—specifically to which actors (public or private, human or technological) they might be assigned, and what constraints or processes might be required in their performance. This analysis can facilitate the evaluation and design of content moderation systems to ensure the capacity and competencies necessary for legitimate, distributed systems of content governance.

Through a combination of methods, legal schemes delegate at least a portion of the responsibility for governing online expression to private actors. Sometimes, statutory schemes assign regulatory tasks explicitly. In others, this delegation often occurs implicitly, with little guidance as to how the treatment of content should be structured. In the law's shadow, online platforms are largely given free rein to configure the governance of expression.

Legal scholarship has surfaced important concerns about the private sector's role in content governance. In response, private platforms engaged in content moderation have adopted structures that mimic public governance forms. Yet, we largely lack the means to

---

DOI: https://doi.org/10.15779/Z383B5W872

† Professor, School of Information, UC Berkeley; Faculty Co-Director, Berkeley Center for Law and Technology, Berkeley, School of Law; Co-Director, Algorithmic Fairness and Opacity Group, School of Information, UC Berkeley. Many thanks to the following persons for thoughtful feedback: organizers David Kaye and Gregory Shaffer, discussant KS Park, and participants at the University of California, Irvine, School of Law Symposium on the Transnational Legal Ordering of Privacy and Speech; moderator David Vladeck and co-panelists Dina Srinivasan and Ashkan Soltani at the National Academies of Sciences, Engineering and Medicine's Committee on Science, Technology, and Law workshop "Section 230 Protections: Can Legal Revisions or Novel Technologies Limit Online Misinformation and Abuse?"; organizers and participants at the 25th Annual BCLT/BTLJ Symposium—Lex Informatica: The Formulation of Information Policy Rules through Technology. Special thanks to Daphne Keller, Nicole Wong, Joris van Hoboken, and Naomi Appelman for providing feedback on subsequent drafts, and to Jessica Li, Khash Goshtasbi, and Sophia Wallach for their detailed feedback, editing, and patience. Research for this article, and the development of the handoff model has been funded by generous support from the US NSF INSPIRE SES1537324.

†† The Rosalinde and Arthur Gilbert Foundation Professor of Law, University of California, Berkeley; Faculty Co-Director, Berkeley Center for Law and Technology.

measure whether these forms are substantive, effectively infusing public values into the content moderation process, or merely symbolic artifice designed to deflect much needed public scrutiny.

This Article's proposed framework addresses that gap in two ways. First, the framework considers together all manner of legal regimes that induce platforms to engage in the function of content moderation. Second, it focuses on the shared set of specific tasks, or subfunctions, involved in the content moderation function across these regimes.

Examining a broad range of content moderation regimes together highlights the existence of distinct common tasks and decision points that together constitute the content moderation function. Focusing on this shared set of subfunctions highlights the different values implicated by each and the way they can be "handed off" to human and technical actors to perform in different ways with varying normative and political implications.

This Article identifies four key content moderation subfunctions: (1) definition of policies, (2) identification of potentially covered content, (3) application of policies to specific cases, and (4) resolution of those cases.

Using these four subfunctions supports a rigorous analysis of how to leverage the capacities and competencies of government and private parties throughout the content moderation process. Such attention also highlights how the exercise of that power can be constrained—either by requiring the use of particular decision-making processes or through limits on the use of automation—in ways that further address normative concerns.

Dissecting the allocation of subfunctions in various content moderation regimes reveals the distinct ethical and political questions that arise in alternate configurations. Specifically, it offers a way to think about four key questions: (1) what values are most at issue regarding each subfunction; (2) which activities might be more appropriate to delegate to particular public or private actors; (3) which constraints must be attached to the delegation of each subfunction; and (4) where can investments in shared content moderation infrastructures support relevant values? The functional framework thus provides a means for both evaluating the symbolic legal forms that firms have constructed in service of content moderation and for designing processes that better reflect public values.

TABLE OF CONTENTS

I. INTRODUCTION

Addressing harms from online content has proven to be a major puzzle of the digital age. It confounds traditional notions of regulation, as First Amendment limits and other non-intervention norms combine with the trans-jurisdictional scope of platform operations to circumscribe the public role in governance of platform content. Through a combination of methods, legal schemes delegate at least a portion of the responsibility for governing expression to private actors. Sometimes, statutory schemes assign regulatory tasks explicitly. Yet this delegation often occurs implicitly, with little guidance as to how the treatment of content should be structured. For example, § 230 of the Communications Decency Act[1] empowers platforms to moderate content by shielding them from liability without specifying processes for, or

1. 47 U.S.C. § 230.

even the content subject to, moderation. In the law's shadow, online platforms are largely given free reign to configure the governance of expression.

Legal scholarship has surfaced important concerns about the private sector's role in content governance. This work includes critiques of particular company processes as opaque, arbitrary, or untethered to important public norms regarding the governance of expression.[2] It also criticizes the use of technology to automate content moderation—a particular form of privatization—for its opacity and for the overbroad, discriminatory, and arbitrary outcomes it produces.[3]

In response to critiques, private platforms engaged in content moderation have adopted structures that mimic public governance forms. One example, Google's Right to be Forgotten Advisory Council involved external experts in the development of rules and processes to guide the private content

2. evelyn douek, *The Limits of International Law in Content Moderation*, 6 U. CALIF. IRVINE J. INT'L., TRANSNATIONAL, & COMPARATIVE L. 37 (2021) ("'[I]nwards-looking, largely public relations-oriented content governance models so widely deployed today' . . . are largely untethered from any particular normative commitments, leaving them unconstrained and arbitrary.").

3. Emma Llansó, Joris van Hoboken, Paddy Leerssen & Jaron Harambam, *Artificial Intelligence, Content Moderation, and Freedom of Expression* (Feb. 26, 2020) (Working Paper, Transatlantic Working Grp., Bellagio Session), https://www.ivir.nl/publicaties/download/AI-Llanso-Van-Hoboken-Feb-2020.pdf; *see* Emma J. Llansó, *No Amount of "AI" in Content Moderation Will Solve Filtering's Prior-Restraint Problem*, 7 BIG DATA & SOC'Y 1 (2020) (arguing that the use of technology to proactively moderate content "acts as a prior restraint on speech, regardless of the accuracy"); Robert Gorwa, Reuben Binns & Christian Katzenbach, *Algorithmic Content Moderation: Technical And Political Challenges In The Automation Of Platform Governance*, 7 BIG DATA & SOC'Y 1 (2020) (arguing algorithmic content moderation increases the opacity of platform practices, exacerbates concerns with fairness and accountability, and obscures important political choices); Hannah Bloch-Wehba, *Automation in Moderation*, 53 CORNELL INT'L L.J. 41 (2020) (arguing ex-ante algorithmic moderation expands platforms' unaccountable control over online speech, exacerbating risks to freedom of speech and association, privacy, and equality). This important literature has produced a wealth of information about the operation of specific instantiations of the content moderation function. It has, moreover, identified and framed important questions implicated by content moderation generally, including the risks of privatizing functions traditionally performed by government entities, the substantive and procedural deficits of platform moderation practices, and the hazards posed by various technical methods of natural language processing used to automate content moderation, such as image recognition and cryptographic hashing. Common themes across this work are concerns with the lack of consistent commitment to normative touchstones, such as international human rights norms, and the lack of attention to context—be it the histories of violence or oppression within particular communities, the events and conversations in which platform content is mobilized, or the divergent meaning and importance of content over time and between communities.

moderation operation.[4] A second example, the most recent and grandiose such structure, is the Facebook Oversight Board (FBOB), frequently referred to as its "Supreme Court."[5] A third example, the most widespread, are "transparency reports," consisting of company-issued public reports disclosing content moderation outcomes along with data about other firm practices that affect the privacy and freedom of expression interests of users.[6]

Longstanding research in "new governance" points to the persistent role of private firms and the need to think constructively, rather than reactively, about the assets and deficits private actors bring to discrete aspects of authority.[7] At the same time, socio-legal scholars warn that governance in private hands can produce symbolic or ceremonial structures that imbue corporate acts with apparent legitimacy but do little to further the public values at stake.[8] Criticism of transparency reporting and the FBOB resonates with these insights.[9]

---

4. Google convened the Council in 2015 in response to the European Union's ruling in *Google Spain v. Agencia Española de Protección de Datos*. Case C131 /12 Google Spain v. Agencia Española de Protección de Datos (AEPD), ECLI:EU:C:2014:317 (May 13, 2014). The Council is no longer operational. *See* Carol A. F. Umhoefer, *Europe: Right to be Forgotten–Google Advisory Council published its report*, Lexology.com (available at https://www.lexology.com/library/detail.aspx?g=db11a725-6a65-4250-b792-9a93d4394057) (noting that the Advisory Council's Final Report was published on February 6, 2015).

5. Facebook convened the Facebook Oversight Board in 2020, and the Board is still in operation today. *See* OVERSIGHT BOARD, https://oversightboard.com/ (last visited Apr. 19, 2022).

6. *See*, *e.g.*, Robert Gorwa & Timothy Garton Ash, *Democratic Transparency in the Platform Society*, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD AND PROSPECTS FOR REFORM 286, 293–299 (Nate Persily & Josh A. Tucker eds., 2020) (discussing major voluntary transparency initiatives of platform companies); Camille François & evelyn douek, *The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting About Information Operations*, 1 J. ONLINE TR. & SAFETY 1, 4–11 (2021) (discussing the history and development of transparency disclosures regarding platforms' actions to address "information operations," an ambiguous category used by platforms to address coordinated, deceptive activity typically targeting clusters of accounts). The three major platforms discussed in this article publish transparency reports to provide the public with some information about content removals, among other actions. *See infra* Section III.A.3.

7. *See infra* note 20.

8. *See* Lauren B. Edelman, *Legal Ambiguity and Symbolic Structures: Organizational Mediation of Civil Rights Law*, 97 AM J. SOC. 1531, 1542 (1992) (exploring the ways that "[l]aws that are ambiguous, procedural in emphasis, and difficult to enforce invite symbolic responses-responses designed to create a visible commitment to law, which may, but do not necessarily," further public values).

9. *See, e.g.*, Monika Zalnieriute, *"Transparency-Washing" In The Digital Age: A Corporate Agenda of Procedural Fetishism*, 8 CRITICAL ANALYSIS L. 39 (2021) (critiquing Transparency Reports of IBM, Google, and Facebook as "procedural fetishism" that provides limited substantive protection).

Rigorous assessment of the "effectiveness" of the structures adopted by platforms requires clarity about the deficits they are meant to address in the private content moderation process. Yet it is often unclear, at a relevant level of specificity, exactly which deficits new structures are targeted to address and against which yardsticks private content moderation systems should be measured. The lack of a clear metric for evaluation stymies efforts to develop and evaluate alternative approaches to structuring and constraining private content moderation and to envision the appropriate role for public law in mandating or catalyzing those alternatives.[10]

This Article suggests two analytic shifts that enable the assessment of content moderation systems' alignment with public values and that enable the design of content moderation systems that ensure the capacity and competencies necessary for legitimate, distributed systems of content governance. Both shifts draw attention to content moderation's function, rather than its form.[11] The first shift involves zooming out—that is,

10. International human rights law has become the dominant framework for assessing platform content governance. *See, e.g.*, David Kaye, *Report of the Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression*, No. A/HRC/38/35 (June 2018) ¶ 41 (arguing for the implementation of "human rights standards transparently and consistently, with meaningful user and civil society input" as a means for holding "both States and companies accountable to users across national borders"); Molly K. Land, *Against Privatized Censorship: Proposals for Responsible Delegation*, 60 VA. J. INT'L L. 363 (2020) (finding that many regulatory regimes that deputize platforms to police and govern content are unlawful under human rights law and proposing a "human rights non-delegation doctrine"); Barrie Sander, *Freedom of Expression in the Age of Online Platforms: The Promise and Pitfalls of a Human Rights-Based Approach to Content Moderation*, 43 FORDHAM INT'L L.J. 939, 966–68 (2020); Evelyn Aswad, *The Future of Freedom of Expression Online*, 17 DUKE L. & TECH. REV. 26, 57–67 (2018). For a discussion of the limits of international human rights law as a tool for reforming content moderation, see douek, *supra* note 2, at 50–64 and Sander, *supra* note 10, at 968–70. Hannah Bloch-Wehba argues that platforms should adopt basic principles of administrative law to ensure accountability to the public. *Global Platform Governance: Private Power in the Shadow of the State*, 72 SMU L. REV. 27 (2019). evelyn douek further notes that "requirements of a clear, precise, and transparent statement of a rule that is justified in the pursuance of a legitimate purpose . . . and due process requirements" are not unique to international human rights or administrative law. douek, *supra* note 2, at 63.

11. Recent work by Niva Elkin-Koren and Maayan Perel points to a different way a focus on functions in the regulation of online expression might be important. *See Separation of Functions for AI: Restraining Speech Regulation by Online Platforms*, 24 LEWIS & CLARK L. REV. 857 (2020). In particular, they recommend distinguishing between "public" functions related to the identification and removal of unlawful content from "private" functions driven by business-related content moderation imperatives, both of which rely on the use of artificial intelligence. These authors suggest that the former should only utilize "independent" AI tools that "embed[ ] public policy." *Id.* at 857–58. In a similar vein, Barrie Sander grounds his human rights analysis by first delineating four platform content moderation activities in which

considering together all manner of legal frameworks that induce platforms to moderate online content. The second involves zooming in—that is, focusing on the shared set of specific tasks, or subfunctions, involved in the content moderation function across these regimes.

Our first analytic shift—zooming out by considering together any legal regimes that structures the governance of online expression—focuses on the functional output rather than the doctrinal basis or problem statement of content moderation. Online content moderation regimes intended to reduce harm arise in a range of contexts. Some, such as those arising under § 230, are widely recognized and labeled as "content moderation" regimes. Others, although they too govern online content, have largely escaped such identification and are generally treated as components of other legal subjects including copyright law, privacy law, human trafficking and child exploitation law, terrorism law, and harassment law.[12]

The wide range of existing statutory schemes that drive the governance of content reflect an extensive variety of approaches .[13] Legal frameworks vary by specificity and tactic—from the detailed regulatory scheme set out in the Digital Millennium Copyright Act (DMCA)[14] to the General Data Protection Regulation's (GDPR) less-specified use of liability risks pursuant to its Right to be Forgotten mandate.[15] Some frameworks directly mandate *what* content should be moderated, such as terrorism-related material, child sexual abuse material,[16] and hate speech. Some specify explicitly *who* should be involved,

---

transparency and oversight should be provided: rulemaking, decision-making, content and advertising, and regulatory compliance. *See* Sander, *supra* note 10 at 998–90.

12. A vast array of other laws can be the source of content removal requests. For example, they can be U.S. Security and Exchange Commission regulations, trade secret law, fraud, or false advertising laws as well as speech that aids or abets illegal conduct (instructional content for criminal or negligent activity).

13. For one taxonomy of approaches states use to enlist private actors in regulating online content, see Molly K. Land, *Against privatized censorship: proposals for responsible delegation*, 60 VA. J. INT'L L. 363, 399 (2019)("[States use] command and control, intermediary liability, and extra-legal influence.")

14. Digital Millennium Copyright Act (DMCA), 17 U.S.C. §§ 512, 1201–05, 1301–32.

15. Regulation (EU) 2016/679, of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons with Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation), art. 17, 2016 O.J. (L 119) 1, 43 (detailing the "Right to erasure ('right to be forgotten')" by providing that "[t]he data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay" where one several grounds applies).

16. Although the term child pornography is still used in legislation in the United States and elsewhere, for our discussion we use the term Child Sexual Abuse Material (CSAM) to

allocating discrete tasks within the content moderation process to a range of public and private actors (i.e., NGOs, content creators, courts, users, and other individuals). Some mandate particular procedural constraints regarding *how* the content moderation process should proceed.

Broadening the lens to include a spectrum of varied content moderation regimes surfaces differences between those regimes. Such analysis highlights the range of approaches to content moderation and offers examples that can be compared to improve content moderation's function.

Yet examining a broader range of content moderation regimes together also points to similarities amongst the different regimes. Specifically, it highlights the existence of distinct common tasks and decision points that together constitute the content moderation function across contexts—tasks and decision points that can each be assigned to different actors and structured differently.

Our second analytic shift—zooming in—focuses on the shared set of tasks, or "subfunctions," involved in the content moderation function across regimes. Content moderation is not a single undifferentiated function. It is comprised of a variety of subfunctions—individual tasks and decisions that may be "handed off" to different human and technical actors to perform in different ways with different normative and political implications. This shift towards a focus on the allocation of responsibility constitutes an application of the "handoff model" for evaluating sociotechnical systems in ethical and political terms developed by Dierdre Mulligan, one of this Article's authors, and information scholar Helen Nissenbaum.[17]

The handoff model focuses on identifying the constituent tasks that comprise systems, here those involving content moderation and the different values implicated by each task. It then considers four questions: What values are important to preserve in structuring those tasks or decisions? Which actors (human or technical), then, should be given the power to perform the task or decision? What procedural or technical constraints should structure the performance of the task or decision? Do the constraints protect or promote the public norms at issue?

This Article uses the handoff model to identify and focus on four key content moderation subfunctions: (1) *definition* of policies, (2) *identification* of potentially covered content, (3) *application* of policies to specific cases, and (4)

---

acknowledge that this material is distinct from adult pornography and depicts child abuse and exploitation.

17. Deirdre K. Mulligan & Helen Nissenbaum, *The Concept of Handoff as a Model for Ethical Analysis and Design*, *in* THE OXFORD HANDBOOK OF ETHICS OF AI 233 (Markus D. Dubber, Frank Pasquale & Sunit Das eds., 2020) [hereinafter *The Concept of Handoff*].

resolution of those cases. This list does not represent an exclusive typology of the stages of content moderation; others might categorize and label subfunctions differently. But these subfunctions reflect the structure of many of the statutes that drive content moderation and reflect insights from the authors' participation in policy processes around the adoption of key legal frameworks driving private platforms' content moderation practices.

Focusing on subfunctions illuminates the ways that the allocation of responsibility to different public or private actors to perform discrete subfunctions can improve legitimacy. Using these four subfunctions supports a rigorous analysis of how to leverage the capacities and competencies of government and private parties throughout the content moderation process. Such attention also highlights how the exercise of that power can be constrained—either by requiring the use of decision-making processes or through limits on the use of automation—in ways that further address normative concerns.

Identifying different subfunctions surfaces three things. First, it illuminates possible choices regarding the allocation of tasks to different actors. As an example, the DMCA assigns the task of identification of potentially infringing material online to content creators.[18] The regime developed to address Child Sexual Abuse Material (CSAM), by contrast, in part relies on machine-learning systems for this subfunction.[19]

Second, focusing on subfunctions makes visible the normative implications of different content moderation configurations. Each identified subfunction implicates different governance norms and demands a different set of competencies. Thus, an assessment of any content moderation regime requires an inquiry into the appropriateness of assigning each subfunction. (For example, which assigned actor has the appropriate incentives, relevant information, or technical capacity in any given context?)

Finally, focusing on discrete subfunctions surfaces constraints, which may take the form of required procedures or limits on automation that can limit different actors' power to perform those tasks. For example, although it may be appropriate to allow a platform to rely on users or a machine-learning system to identify material for review under a content moderation policy, allowing reliance on either to define the content subject to moderation may raise concerns.

Thus, this Article's functional framework generates constructive insights that are both concrete and generalizable across contexts. It offers a means for

18. *See* 17 U.S.C. § 512 (establishing the notice and takedown procedure).

19. *See infra* text accompanying notes 130–139.

critically assessing the way various content moderation regimes both allocate and constrain various subfunctions. It identifies the implications of those different arrangements for public values. And it considers how these subfunctions might be appropriately structured to forge legitimate content governance systems going forward. This Article proceeds in the following way.

Part I explores three structures that private platforms have adopted in response to legitimacy critiques: Google's Advisory Council on the Right to be Forgotten, Facebook's Oversight Board, and the use of transparency reports. Noting the shortcomings of existing frameworks to assess the effectiveness and appropriateness of these regimes, this Part establishes the need for an evaluative framework that focuses on the subfunction deficits of the content moderation process that each structure attempts to address.

Part II sets forth this new evaluative framework. It first identifies and discusses four key subfunctions, the public values each implicates, and the competencies each requires. It then explores case studies from diverse content moderation regimes to illustrate possible task allocations to ensure that the appropriate competencies are brought to bear. Examples of constraints that in different circumstances have been—or could be—imposed on the performance of each subfunction include limits both on the process and on the identity of the actors (human and technical) by whom subfunctions may be implemented. Thus, the framework considers ways that largely private decision systems can best be held accountable to key normative values.

This framework reflects the insights of our and others' work in "new governance" scholarship, emphasizing the importance of identifying, surfacing, structuring, and constraining decision-making delegated to private actors in governance in light of accountability to public norms.[20] By engaging in a precise exploration of how each actor enlisted in a content moderation regime executes the function delegated to it, the framework allows consideration of the ways in which replacing one actor with another disrupts (or doesn't disrupt) the ethical and political dimensions of the subfunction and the configuration of values in the system as a whole.

20. *See, e.g.*, Kenneth A. Bamberger, *Technologies of Compliance: Risk and Regulation in a Digital Age*, 88 TEX. L. REV. 669, 684 (2010) [hereinafter Bamberger, *Technologies of Compliance*] (focusing on the "decisionmaking processes of private actors" acting "as partners in regulation"); Kenneth A. Bamberger, *Regulation as Delegation: Private Firms, Decisionmaking, and Accountability in the Administrative State*, 56 DUKE L.J. 377 *passim* (2006) [hereinafter Bamberger, *Regulation as Delegation*]; Kenneth A. Bamberger & Deirdre K. Mulligan, *New Governance, Chief Privacy Officers, and the Corporate Management of Information Privacy in the United States: An Initial Inquiry*, 33 L. & POL'Y 477, 480–82 (2011) (summarizing the "new governance" literature).

Finally, Part III applies this analytic framework. First, it evaluates to what extent (if at all) the symbolic content moderation forms discussed in Part II provide competencies that address the relevant democratic deficiencies in the subfunction at issue. Second, making a constructive turn, it uses the Global Internet Forum to Counter Terrorism Shared Industry Hash Database[21] to illustrate how to allocate subfunctions and coordinate subfunction constraints. As an example, it applies the framework to the Global Internet Forum to Counter Terrorism Shared Industry Hash Database. It then compares that content moderation configuration to the National Center for Missing and Exploited Children's Hash Database to highlight the different characteristics of regulated content and the different allocations of responsibility for the databases. This comparison illustrates the ways that the choice about which actor should be given authority for those databases affects the extent to which investments in shared content moderation infrastructures can support public values, including transparency, legitimacy, nondiscrimination, rational decision-making, and the promotion of competition.

Dissecting the allocation of subfunctions in various content moderation regimes reveals the distinct ethical and political questions that arise in alternate configurations. Specifically, it offers a way to think about four key questions: (1) what values are most at issue regarding each subfunction; (2) which activities might be more appropriate to delegate to particular public or private actors; (3) which constraints need to attach to the delegation of each subfunctions; and (4) where investments in shared content moderation infrastructures could support relevant values? The functional framework thus provides a means for evaluating the symbolic legal forms that firms have constructed in service of content moderation.

This Article's functional framework applies the handoff model to the content moderation landscape. Too often the salient differences in a new or competing functional arrangement of content moderation comes to light only after adoption. A functional framework offers a means to frontload this values analysis, allowing regulators, system designers, and other stakeholders to foresee, at least to some extent, and prioritize values during design. Looking forward, this Article's proposed framework enables analyses of how the mix of actors used to perform a function matters even before a content moderation regime is put into place.

21. *Tech Innovation*, GLOB. INTERNET F. TO COUNTER TERRORISM, https://gifct.org/tech-innovation (describing the hash-sharing database).

II. SHIFTING THE FOCUS FROM FORM TO FUNCTION

This Part first describes the structures platforms have adopted to insulate their private content moderation practices from critiques claiming they fail to sufficiently satisfy public governance values. It next unpacks the specific content moderation *subfunction* that each structure seeks to legitimate and, with this insight, suggests a new analytic direction for evaluating the alignment of content moderation systems with public values.

A. LEGALISTIC FORM AS A RESPONSE TO LEGITIMACY CRITIQUES

Faced with critiques questioning the legitimacy of their private content moderation activities,[22] social media platforms have attempted to reshape the roles and responsibilities for moderating content online by evolving what social and technical studies (STS) theorists would term content management "scripts."[23] Most visibly, these scripts include what legal sociologist Lauren Edelman refers to as "symbolic legal structures"—organizational approaches that evoke a notion of legality.[24] Through the adoption of structures that mimic those of public legal institutions—such as reports, public hearings, and appellate review structures—platforms seek to convey a sense of legitimacy to actions that are frequently entirely unrelated to public legal mandates.

This Section describes three paradigmatic examples of symbolic legal structures:

(1) Google's Advisory Council on the Right to be Forgotten (GAC)

(2) Facebook's Oversight Board (FBOB)

(3) Transparency Reports (TR)[25]

These structures mimic legal institutions and claim attributes often associated with public governance: government advisory committees/ expertise and stakeholder participation (GAC); courts/expertise and independence (FBOB); and public access to the outcomes of adversary

22. *See supra* notes 2–3.

23. MADELEINE AKRICH, THE DE-SCRIPTION OF TECHNICAL OBJECTS, *in* SHAPING TECHNOLOGY/BUILDING SOCIETY: STUDIES IN SOCIOTECHNICAL CHANGE 205 (Wiebe E. Bijker & John Law eds., 1992).

24. LAUREN B. EDELMAN, WORKING LAW: COURTS, CORPORATIONS, AND SYMBOLIC CIVIL RIGHTS 101–02 (2016).

25. Platforms use Transparency Reports to publicly disclose data about the outcomes and legal bases of content removal requests they receive from different jurisdictions around the world, and, in a few instances, some information about removals pursuant to platform policies.

processes, particularly those where the government is a party/transparency (TR).

1. Google's Advisory Council on the Right to be Forgotten

Google's Advisory Council on the Right to be Forgotten arose in response to the European Court of Justice's 2014 decision in *Google Spain SL v. Agencia Española de Protección de Datos* (hereinafter *Google Spain*).[26] The ruling clarified that Google, as a data controller under data protection law, had an independent obligation to comply with the General Data Protection Regulation, Europe's data protection law. The court reasoned that because the information was no longer necessary for the purpose of real-estate auction, the plaintiff's interest in privacy overrode the "interest of the general public" in having access to private information. Therefore, the information had to be removed from Google's search results.

The holding left Google in the difficult position of having to develop decision criteria and processes for handling privacy objections to search engine results. Specifically, the ruling required Google to evaluate going forward whether an individual has a cognizable data protection interest and weigh that interest against "the preponderant interest of the general public in having . . . access to the information."[27] Thus Google was thrust into definitional work—policy formation—not just implementation.

The company objected to the role the court asked it to play throughout the *Google Spain* dispute, and many advocates and scholars shared Google's concerns. However, the concerns about the assignment to private actors of the authority to engage in such definitional work are not unique to the right to be forgotten.

---

26. Case C131 /12 Google Spain v. Agencia Española de Protección de Datos (AEPD) (Google Spain), ECLI:EU:C:2014:317 (May 13, 2014), http://curia.europa.eu/juris/document/document.jsf?text=&docid=152065&doclang=EN (finding Google had an obligation to delist search results that would normally be returned in response to queries on an individual's name where those results interfered with the privacy of the individual). *Id.* The E.U. Data Protection Directive developed guidance through various working groups to help companies comply with the European General Data Protection Regulation (GDPR). The GDPR, which came into force on May 25, 2018, adopted the right to be forgotten framework developed in the European Court of Justice's decision. 2016 O.J. (L 119) 1.

27. While this is often discussed as a privacy interest, it is more specifically a data protection interest arising under art. 8 rather than art. 7 of the Charter of Fundamental Rights of the European Union, 2012 O.J. (C 326) 2. For a useful discussion of the distinction and its implications for content removals by online service providers, see Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation*, 33 BERKELEY TECH. L. J. 287, 315–18 (2018).

Under many other legal frameworks, platforms are responsible for determining which content is subject to removal under the relevant policies (public or private). Liability regimes that put platforms at risk but give them no, or at least limited, guidance about how to apply broad legal standards delegate significant definitional work to platforms, like the *Google Spain* decision that led to the creation of the GAC. Moreover, legal frameworks like Section 230, which shield content moderation activities of social media platforms irrespective of the policies or processes they use to do it, delegate nearly all definitional work to platforms.

Faced with a regulatory system that delegates responsibility for developing both substantive decision-making criteria and processes for enforcing public law (rather than platform policy), Google employed a structure that resembled those used by government entities to develop policy: the government advisory committee. Google appointed an advisory council to design a system to process removal requests and to provide substantive decision-making criteria regarding how to balance an individual's right to privacy with the public's interest in access to information in the moderation of content implicated by the right to be forgotten.[28]

In so doing, the company chose to adopt a form that resembled a common symbolic legal structure that government bodies use to garner advice about complex and often contentious policy choices. From federal advisory committees to congressionally created commissions, government entities frequently enlist outside experts in the assessment of substantive policy matters to ensure that expert knowledge and the perspectives of multiple constituencies are included and that the committee's work is relatively transparent and deliberative. For example, the Federal Advisory Committee Act (FACA),[29] passed in 1972, both (1) provides for the involvement of committees composed of experts, representatives of stakeholders, and representatives with different political views as tools for providing policy advice to U.S. government entities, and (2) imposes constraints on the composition of committees requiring them to be "fairly balanced in terms of the points of view represented and the functions to be performed" and have enough autonomy from the appointing power (Congress, the President, or an

28. *See generally* LUCIANO FLORIDI, SYLVIE KAUFFMAN, LIDIA KOLUCKA-ZUK, FRANK LA RUE, SABINE LEUTHEUSSER-SCHNARRENBERGER, JOSÉ-LUIS PIÑAR, PEGGY VALCKE & JIMMY WALES, REPORT TO THE ADVISORY COUNCIL TO GOOGLE ON THE RIGHT TO BE FORGOTTEN (2015).

29. Federal Advisory Committee Act, 5. U.S.C. § 2(a) (stating "they are frequently a useful and beneficial means of furnishing expert advice, ideas, and diverse opinions to the Federal Government" but imposing conditions).

agency head) to limit undue influence.[30] These constraints ensure FACA committees are independent and representative of various stakeholders. FACA also requires transparency in the information committees rely on and their decision-making processes. Most committee meetings must be noticed in the Federal Register and open to the public,[31] and committee materials must be made available for public inspection.[32]

Google's Advisory Council, in turn, consisted of eight members[33] selected by the company but with "no contractual relationship with Google on this project."[34] To further signal the council's independence from Google, Google did not require the GAC members to sign non-disclosure agreements and only reimbursed them for travel costs associated with the public and private meetings necessary for the project.[35] To support the work, the GAC reportedly relied on a range of documents, including non-confidential "publicly available" information, European Court of Human Rights case law, policy guidelines of news organizations, and the Article 29 Working Party's Implementation Guidelines. The GAC solicited expertise from Google and outside experts. Google provided the GAC with briefings from three experts: "an engineer, who explained Search; a Google lawyer,[36] who explained their compliance procedures; and a lawyer from an outside law firm, who explained the legal basis of the Ruling."[37] To gather additional expert opinions and stakeholder input, the GAC held seven public consultations with experts in Europe and gathered expert and lay opinions through a website.[38] In addition to these

30. 5 U.S.C. § 2(b).

31. 5 U.S.C. § 10(a)(2).

32. 5 U.S.C. § 11.

33. The eight members are Luciano Floridi, Professor of Philosophy and Ethics of Information at the University of Oxford; Sylvie Kauffman, Editorial Director, Le Monde; Lidia Kolucka-Zuk, Director of the Trust for Civil Society in Central and Eastern Europe; Frank La Rue, UN Special Rapporteur on the Promotion and Protection of the Right to Freedom of Opinion and Expression; Sabine Leutheusser-Schnarrenberger, former Federal Minister of Justice in Germany; José-Luis Piñar, Professor of Law at Universidad CEU and former Director of the Spanish Data Protection Agency (AEPD); Peggy Valcke, Professor of Law at University of Leuven; Jimmy Wales, Founder and Chair Emeritus, Board of Trustees, Wikimedia Foundation. *Read the Advisory Council's final report*, GOOGLE ADVISORY COUNCIL, https://archive.google.com/advisorycouncil.

34. Jean-Marie Chenou & Roxana Radu, *The "Right to Be Forgotten": Negotiating Public and Private Ordering in the European Union*, 58 BUS. & SOC'Y 74, 90 (2019).

35. *Id.*

36. Daphne Keller was then Associate General Counsel for Intermediary Liability at Google and is now Director of Intermediary Liability at Stanford Law School's Center for Internet and Society.

37. FLORIDI ET AL., *supra* note 28, at 2. It is unclear whether the engineer was a Google employee or not. Google also provided staff support. *Id.*

38. *Id.* at 1-2.

public consultations, the GAC held three council-only meetings in which they deliberated and formulated guidance.[39] The GAC's final report, published on February 6, 2015, discussed alternative proposals that testifying experts offered at the seven public consultations; however, it did not reference the public comments received.[40] There is no public record of those comments.[41]

The GAC final report framed its guidance as addressing the right to "delisting"—the removal of "links returned in search results based on an individual's name when those results are '*inadequate, irrelevant or no longer relevant, or excessive.*"'[42] This was required under the *Google Spain* decision interpreting search engines obligations under Article 14 of the EU Data Protection Directive and the exceptions to this right where there is an overriding public interest in results "for particular reasons, such as the role played by the data subject in public life."[43] Of particular importance to the report's guidance is the conclusion that "whether the data subject experiences harm" from inclusion in a name-based search results page is "relevant to [the] balancing test" required to determine exceptions to the delisting right.[44] This conclusion appears to be at odds with the *Google Spain* ruling which makes no mention of a harm assessment and states, "it is not necessary in order to find such a right [to delisting] that the inclusion of the information in question in the list of results causes prejudice to the data subject."[45] To inform its understanding of harm,[46] the GAC drew on case law addressing rights to data protection, privacy, and freedom of expression and information from the Court of Justice for the European Union (CJEU) interpreting the Charter of Fundamental Rights of the European Union[47] and the European Court of Human Rights (ECHR) interpreting the European Convention on Human Rights.[48] Based on the GAC's analysis, the report concluded that "[t]he ruling, while reinforcing European citizens' data protection rights, should not be interpreted as a legitimation for practices of

39. *Id.*

40. *Id.* at 34–37.

41. Chenou & Radu, *supra* note 34, at 88 ("[T]he comments submitted in response to the Request for Comments form on the Advisory Council's website have not been published.").

42. FLORIDI ET AL., *supra* note 28, at 2 (citing *Google Spain*, at ECLI:EU:C:2014:317, ¶ 94).

43. *Google Spain*, at ECLI:EU:C:2014:317, ¶ 97.

44. FLORIDI ET AL., *supra* note 28, at 6.

45. *Google Spain*, at ECLI:EU:C:2014:317, ¶ 96.

46. *Id.* at ¶ 97.

47. The Charter of Fundamental Rights of the European Union establishes the right to privacy (article 7) the right to data protection (article 8) and the right to freedom of expression and information (article 11). 55 O.J. (C 326) 391, 397–98.

48. The European Convention on Human Rights establishes the right to privacy (article 8) and freedom of expression (article 10). European Convention on Human Rights arts. 8, 10, Nov. 4, 1950 (amended 1998).

censorship of past information and limiting the right to access information."[49]

The introduction of harm assessment restructures the test set out in the *Google Spain* ruling from one that looks at exceptions to data subjects' rights necessitated by the "preponderant interest of the general public" (an inquiry looking at specific facts, such as the subject's role in public life) to one that simply contrasts the harms to data subjects against the "preponderant interest of the general public."[50] The GAC recommended that Google consider four key criteria when evaluating delisting requests: (1) the data subjects role in public life; (2) the nature of the information; (3) the source of the information, including its motivation for publication; and (4) time, "the notion that information may at one point be relevant, but as circumstances change, the relevance of that information may fade."[51] While the criteria and sub-criteria set out by the GAC capture many of the key concepts set out by the Article 19 Data Protection Working Party in their guidance, the introduction of balancing harms to data subjects against the public interest could substantially alter the outcomes achieved despite those shared criteria.

2. Facebook's Oversight Board

The most recent symbolic legal structure to emerge is Facebook's Oversight Board, a body explicitly likened to the most venerated of legal structures—the Supreme Court. In a trial balloon of the concept, Mark Zuckerberg said, "You can imagine some sort of structure, almost like a Supreme Court, that is made up of independent folks who don't work for Facebook, who ultimately make the final judgment call on what should be acceptable speech in a community that reflects the social norms and values of

49. FLORIDI ET AL., *supra* note 28, at 6.

50. *Id.* at 5–6 (citing *Google Spain*, at ECLI:EU:C:2014:317, ¶ 97).

51. *Id.* at 7–14. The report went beyond the mandate to advise Google on decision-making criteria and included recommendations on the processes and inputs for decision-making. *Id.* at 15–21. It also included a discussion of the alternative ideas and technical proposals to establish adjudication processes presented during public consultations. *Id.* at 34–37. The alternative proposals discussed attend to democratic deficits inherent to the regulatory framework—as opposed to those that Google could address through implementation choices—including the lack of an administrative appeals process for users whose content was removed and the reliance on private rather than public processes to adjudicate claims. Among other proposals, experts suggested establishing "a clear channel of appeal to a public authority for publishers seeking vindication of Article 10 rights, parallel to data subjects' right of appeal to DPAs," and "a public mediation model, in which an independent arbitration body assesses removal requests," modeled on the domain name dispute resolution process. *Id.* at 36.

people all around the world."[52] Zuckerberg explained that he had "come to believe that Facebook should not make so many important decisions about free expression and safety on [its] own."[53]

Facebook established the FBOB after receiving mounting criticism from policymakers, journalists, and the public regarding its decisions to remove and maintain content. As described above, many legal frameworks delegate substantial *definitional* work to platforms. In addition, many legal frameworks delegate responsibility for the *application* of the rules, and the processes and tools used to do so, to private platforms. The court analogy Zuckerberg chose responded to specific concerns with the *application* subfunction of the content moderation task.

In November 2018, Mark Zuckerberg announced a "blueprint" for increasing transparency and accuracy in content removals.[54] He outlined a number of internal changes, including improving Facebook's efforts to independently identify and remove content in violation of Facebook's community standards rather than relying on reports from users. However, the most significant change Zuckerberg announced was the creation of an independent oversight board to handle at least some content removal appeals.[55] To solicit guidance on the creation and function of the board and develop stakeholder buy-in, Facebook undertook an extensive consultation process involving public for a,[56] smaller expert working groups, as well as town halls.[57]

52. Ezra Klein, *Mark Zuckerberg on Facebook's hardest year, and what comes next*, VOX (Apr. 2, 2018, 6:00 AM EST), https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge.

53. Mark Zuckerberg, *A Blueprint for Content Governance and Enforcement*, FACEBOOK (Nov. 15, 2018), https://www.facebook.com/notes/mark-zuckerberg/a-blueprint-for-content-governance-and-enforcement/10156443129621634.

54. *Id.* Peter Stern, Head of Product Policy Stakeholder Engagement, has indicated that the removal of the "After the 'Terror of War' " and the ensuing controversy, was the impetus for these sweeping reforms. *See* Kate Klonick, *Facebook v. Sullivan*, KNIGHT FIRST AMEND. INST. (Oct. 1, 2018), https://knightcolumbia.org/content/facebook-v-sullivan (reporting on an interview with Peter Stern).

55. Kate Klonick exhaustively documented the path to the creation of the Board. *See generally The Facebook Oversight Board: Creating an Independent Institution to Adjudicate Online Free Expression*, 129 YALE L. J. 2418 (2020).

56. *See* Brent Harris, *Getting Input on an Oversight Board*, META (Apr. 1, 2019), https://about.fb.com/news/2019/04/input-on-an-oversight-board/ (describing public consultation process); Klonick, *supra* note 52, at 2448–57 (describing the consultation process as well as internal processes around the creation of the Board).

57. *See* Brent Harris, *Global Feedback and Input on the Facebook Oversight Board for Content Decisions*, META (June 27, 2019), https://about.fb.com/news/2019/06/global-feedback-on-oversight-board/ (describing an entire process, which included expert consultations and public

The creation of the FBOB was Facebook's effort to address the largely unchecked discretion and lack of transparency around processes used to enforce its own community standards. Facebook established the FBOB under the shadow of Section 230. Pursuant to the statute's content moderation regime, platforms need not notify those whose content is subject to moderation; they need not respond to requests for moderation; they need not provide the reasons for moderation decisions; they need not provide any means to challenge content moderation activities; and they need not provide information about the processes, tools, and rules (if they exist) that guide their decision-making.

The wide latitude that platforms like Facebook have to moderate content under Section 230 permitted a lack of transparency about the rules and their application that were criticized as lacking the hallmarks of substantive and procedural legitimacy associated with adjudicating disputes in the public sector. Neither the parties nor the public were privy to the rules Facebook applied, the kinds of individuals or technology tasked with the application, or the controlling processes. Societal stakeholders raised concerns about Facebook's lack of independence, how various incentives might influence their application of rules, and the lack of processes to contest content moderation decisions.[58]

In establishing the FBOB, Facebook emulated the most symbolic of legal structures. The FBOB borrows features that are evocative of courts. The nine-page charter establishing the board contains detailed sections about membership, scope of authority, board procedures, implementation, board governance amendments and bylaws, and compliance with law.[59] Both Facebook and users can appeal to the FBOB for review of a content moderation decision, but the board retains discretion over which cases to hear. The charter establishes that the FBOB's task is to consider whether decisions were "consistent with Facebook's content policies and values" and that FBOB decisions set precedent. In the charter, Facebook committed to upholding the

consultation, and releasing report on the construction of the oversight board); Klonick, *supra* note 53, at 2448–57 (describing the consultation process as well as internal processes around the creation of the Board).

58. *See, e.g.*, *The Santa Clara Principles On Transparency and Accountability in Content Moderation*, SANTA CLARA 1.0, https://santaclaraprinciples.org (last visited Mar. 18, 2022) (operational principle three calling on platforms to online speech platforms to create "meaningful opportunity for timely appeal" of moderation decisions and consider establishing "independent external review processes").

59. *Oversight Board Charter*, FACEBOOK, https://about.fb.com/wp-content/uploads/2019/09/oversight_board_charter.pdf (last visited Mar. 18, 2022) [hereinafter *Oversight Board Charter*].

rulings of the Board as final unless it would violate the law to do so.[60] The charter also contains several provisions to bolster the FBOB's independence.[61] In May of 2020, Facebook announced the first members of the Board.[62] Board members included individuals with "experience in press freedom, digital rights, religious freedom, content moderation, online safety, internet censorship, platform transparency and technology."[63]

Together, the lack of transparency into how platforms apply rules to specific content, the absence of notice and participation rights for relevant parties, and the lack of a public record of the reasoning behind determinations have undermined the perceived legitimacy of platforms' application of content moderation standards. The court-like aspects of the FBOB evince a direct attempt to gain legitimacy by adopting structures and processes that support the independence of decision makers, transparency of decisions in the particular case, and affected parties' participation in public adjudicatory processes.

3. Transparency Reports

Over the last twelve years, social media platforms and other information and communication technology companies have begun releasing "transparency reports," which share information about content removals and disclosures of users' personal information.[64] Google released the first transparency report in 2010, and many other technology companies have adopted some version of transparency reporting in the intervening years.[65]

60. Catalina Botero-Marino, Jamal Greene, Michael W. McConnell & Helle Thorning-Schmidt, *We Are a New Board Overseeing Facebook. Here's What We'll Decide*, N.Y. TIMES (May 6, 2020), https://www.nytimes.com/2020/05/06/opinion/facebook-oversight-board.html.

61. The Board is funded through an independent trust, which was set up by Facebook but cannot be revoked by the company. Each board member will serve fixed terms of three years and may serve up to three terms. *Announcing the First Members of the Oversight Board*, OVERSIGHT BD. (May 6, 2020), https://www.oversightboard.com/news/announcing-the-first-members-of-the-oversight-board/.

62. Nick Clegg, *Welcoming the Oversight Board*, META (May 6, 2020), https://about.fb.com/news/2020/05/welcoming-the-oversight-board/. In August of 2019, Facebook had issued further guidance, explaining how the Board would operate. *Facebook Oversight Board for Content Decisions: What to Know*, META (Aug. 22, 2019), https://www.facebook.com/journalismproject/facebook-oversight-board-for-content-decisions-overview.

63. *Id.*

64. Peter Micek & Isedua Oribhabor, *The what, why, and who of transparency reporting*, ACCESS NOW (Apr. 2, 2020), https://www.accessnow.org/the-what-why-and-who-of-transparency-reporting/.

65. *See* James Losey, *Surveillance of Communications: A Legitimization Crisis and the Need For Transparency*, 9 INT'L J. COMMC'N 3450, 3453 tbl.1 (2015) (providing an overview of Transparency Reports of forty-one information and communication technology sector

Companies initially adopted transparency reports to increase public visibility into government requests for the personal information and communications of platform users [66] and, for some platforms, content removals made at the behest of third parties.[67] The reports adopted the form of wiretap reports, which the U.S. Department of Justice files yearly detailing the number and categories of state and federal wiretaps requested and issued.[68] By mirroring a practice used in public governance, companies sought to "reassure national and international subscribers that the[y] had rigorous processes for evaluating government requests for data, shed light on the

companies, documenting that some publish data about removals in multiple countries, all provide data about demands for personal data, but not all publish data about content removal, and to the extent they do the data varies in coverage); *Transparency Reporting Index*, ACCESS NOW, https://www.accessnow.org/transparency-reporting-index/ (last visited Apr. 20, 2022). The earliest effort to provide transparency about content removed by social media platforms was the Chillingeffects.org website. Maintained by a set of law school clinics, the website was a repository for DMCA takedown notices. In 2003, Google began contributing the takedown notices and, importantly, providing a link to the site where relevant results had been removed. A few smaller internet service providers contributed takedown notices as well. However, more recently some companies have "quietly dropped the practice," and "not a single household-name tech firm seems to have adopted [the reports] since early 2016." Rob Pegoraro, *Tech Companies Are Quietly Phasing Out a Major Privacy Safeguard*, THE ATLANTIC (Sept. 29, 2019), https://www.theatlantic.com/technology/archive/2019/09/what-happened-transparency-reports/599035/.

66. The first Ranking Digital Rights Corporate Accountability Index published in 2015 shows that the sixteen companies evaluated were generally providing more transparency into government requests for customer data than government requests to remove content. *Compare P11. Data about third-party requests for user information*, RANKING DIGIT. RTS., https://rankingdigitalrights.org/index2015/indicators/p11/ (last visited Apr. 20, 2022), *with F7. Data about Government Requests*, RANKING DIGIT. RTS., https://rankingdigitalrights.org/index2015/indicators/f7/ (last visited Apr. 20, 2022).

67. Google provides a brief history of the development of their transparency reports at GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/about?hl=en (last visited Apr. 20, 2022). From inception, they were designed to provide information about both demands for user data and government requests for content removal. In 2009, Google had been blocked in 25 different countries but wanted to reveal that short of full blocking governments were demanding the removal of specific content. Nicole Wong, Dinner Speech at Conference on Liberation Technology in Authoritarian Regimes 4 (Oct. 11, 2010), https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/evnts/media/2010-10_Nicole_Wong_Stanford_Liberation_Technology.pdf ("[W]e wanted to bring some transparency to what is certainly only our limited view on government activity on the Internet. It is not complete. It is not complete because it is only about our products. It is not complete because our data is not sufficiently granular enough yet. It is not complete because some governments will not even let us publish this data. So, why did we do it? The conversation about government censorship and surveillance has to start somewhere.").

68. *See, e.g.*, *Wiretap Report*, U.S. CTS., https://www.uscourts.gov/statistics-reports/wiretap-report-2019 (Dec. 31, 2019) (providing information on wiretaps by jurisdiction and crime, as well as other information).

regularity and breadth of such requests, and encourage reforms in government surveillance activities."[69]

Since 2010, the use and scope of transparency reports have expanded in two ways. First, they now provide information about content removals in response to legal processes by private parties as well as government entities,[70] shedding light on what legal scholar Jack Balkin has called "new-school speech regulation," by which owners of platforms and other digital infrastructure are "coerce[d] or co-opt[ed]" into regulating speech.[71]

Second, in response to public criticism of content removals under platforms' Terms of Services,[72] the Reports increasingly include information about

69. Christopher Parsons, *The (in) Effectiveness of Voluntarily Produced Transparency Reports*, 58 BUS. & SOC'Y 103, 112 (2019); *see, e.g.*, *Twitter Transparency Center*, TWITTER, https://transparency.twitter.com/en/about.html (last visited Mar. 18, 2022) ("The original goal of our transparency report was to provide the public with recurring insights into government pressures that impacted the public.").

70. Beginning in 2012 Google began providing data about content removals pursuant to DMCA requests. As Daphne Keller and Paddy Leersen discuss, current transparency reports do not provide data about removal requests under all laws. DAPHNE KELLER & PADDY LEERSEN, FACTS AND WHERE TO FIND THEM: EMPIRICAL RESEARCH ON INTERNET PLATFORMS AND CONTENT MODERATION, *in* SOCIAL MEDIA AND DEMOCRACY: THE STATE OF THE FIELD, PROSPECTS FOR REFORM 220, 228 (N. Persily, & J. A. Tucker eds., 2020) ("[M]ost transparency reports only cover particular categories of takedowns—often only those initiated by governments or copyright-holders. This leaves open questions about platforms' responses to legal allegations brought by individuals under, say, French defamation law or Brazilian privacy law.").

71. Jack M. Balkin, *Free Speech is a Triangle*, 118 COLUM. L. REV. 2011, 2016–17 (2018) (describing risks of new-school speech regulation: collateral censorship and digital prior restraint). This effort was also informed by an earlier effort to highlight the behavior of copyright holders wielding the power of the DMCA: the Chilling Effects database founded in 2002. LUMEN, https://lumendatabase.org/ (last visited Mar. 18, 2022). That database, a collaborative project established by a group of law school clinics and the Electronic Frontier Foundation, collected copies of takedown requests that were contributed by recipients, Google, and some smaller service providers; annotated and archived them; and made them available for retrieval through search engines. Google, an initial contributor of notices, included a notice about search results removed due to DMCA requests at the bottom of their search results page and provided a link to the relevant takedown request in the Chilling Effects database. This infrastructure provided web searchers and researchers insight into information removed in response to DMCA requests. Chilling Effects was replaced by Lumen and is run by the Berkman-Klein Center for Internet and Society. *About Us*, LUMEN, https://lumendatabase.org/pages/about (last visited Apr. 20, 2022).

72. These include Facebook's removal of a Pulitzer Prize winning graphic photo of a naked Vietnamese girl suffering as napalm from a U.S. attack burned her skin (The Terror of War), ongoing concerns about biases of all sorts in corporate content moderation, and the government use of terms of service violations as a quick and invisible way to remove content, including content they may be unable to remove through legal processes. Zoe Kleinman, *Fury over Facebook 'Napalm girl' censorship*, BBC NEWS (Sept. 9, 2016), https://www.bbc.com/news/

content removals under their terms of service. In 2015, Ranking Digital Rights, which annually evaluates company policies and practices affecting speech and privacy, reported that no company was regularly publishing data about content moderation or account suspensions based on the company's terms of service or other rules.[73] By 2019, for example, Facebook, Google, and Twitter disclosed "comprehensive data about content removals due to terms of service enforcement" and Microsoft was publishing some information, although less comprehensive, about terms of service enforcement.[74]

Today, social media platforms use transparency reports to provide data on the extent and kind of government queries for user data (privacy); the extent, kind, and action platforms take in response to requests for content removal by government and private parties; and content removals under their terms of service. Twitter's Transparency Center—an interactive online version of its original transparency reports—reflects this evolution. The Center now provides relatively robust data and visualizations about the company's content moderation activities.[75] Twitter includes data about external requests for data and internal content moderation activities taken in response to a wide range of laws as well as Twitter's rules. The Center provides jurisdiction-specific data, an interactive tool for comparing countries against each other, data about trends of across time and jurisdictions, and examples offering detail and

technology-37318031; Espen Egil Hansen & Dear Mark, *I Am Writing This to Inform You that I Shall Not Comply with Your Requirement to Remove This Picture*, AFTENPOSTEN (Sept. 8, 2016, 9:33 PM), https://www.aftenposten.no/meninger/kommentar/i/G892Q/dear-mark-i-am-writing-this-to-inform-you-that-i-shall-not-comply-wit; Michael Nunez, *Former Facebook Workers: We Routinely Suppressed Conservative News*, GIZMODO (May, 9, 2016), https://gizmodo.com/former-facebook-workers-we-routinely-suppressed-conser-1775461006; *see* Brian Chang, *From Internet Referral Units to International Agreements; Censorship of the Internet by the UK and EU*, 49 COLUM. HUM. RTS. L. REV. 114 (2017).

73. RANKING DIGIT. RTS., 2019 RDR CORPORATE ACCOUNTABILITY INDEX 42 (2019) [hereinafter RDR 2019], https://rankingdigitalrights.org/index2019/assets/static/download/RDRindex2019report.pdf.

74. *Id.* Twitter was the first to disclose information about actions taken under its Terms of Service in its transparency report. RDR 2019 also found that all companies reviewed disclosed basic information about their terms of service. *Id.* at 44. The most recent version of Ranking Digital Rights' Corporate Accountability Index, reviewing twenty-six companies, can be found at *The 2020 RDR Index*, 2020 RANKING DIGIT. RTS. CORP. ACCOUNTABILITY INDEX, https://rankingdigitalrights.org/index2020/ (last visited Apr. 20, 2022). Google maintains a non-exhaustive list of entities publishing transparency reports at GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/ (last visited Apr. 20, 2022).

75. *Twitter, Inc.*, 2020 RANKING DIGIT. RIGHTS CORP. ACCOUNTABILITY INDEX, https://rankingdigitalrights.org/index2020/companies/Twitter (last visited Apr. 19, 2022) (finding that Twitter shared more data about the enforcement of platform rules than its peers and that it discloses more data about government demands for content removal and user information than most of its U.S. peers; ranking Twitter #1, with a score of 53% on the Index).

context about specific actions.[76] Yet, despite the steady expansion, transparency reports do not provide a complete picture of content removals or content moderation practices[77] or requests for personal information.[78]

B. SHIFTING THE FOCUS FROM FORM TO FUNCTION

By engaging the public, bringing in outside experts, and producing records of decision-making in forms that resemble those of traditional legal institutions, platforms have attempted to legitimize substantive outcomes through the adoption of processes associated with the legitimate institutional exercise of legal power. Yet commentators have expressed a deep sense that "[t]he inwards-looking, largely public relations-oriented content governance models so widely deployed today are unsatisfying."[79]

---

76. *See, e.g.*, *Removal Requests*, TWITTER TRANSPARENCY (Dec. 2020), https://transparency.twitter.com/en/reports/removal-requests.html#2020-jul-dec (finding that through the "removal requests" report section a visitor to the Transparency Center can view worldwide statistics and a short analysis section noting that Japan accounts for 43% of global requests and that those requests are primarily related to laws regulating control substances, obscenity, and money lending.) The visitor could then select the Japan specific report and review more data. *Japan*, TWITTER TRANSPARENCY, https://transparency.twitter.com/en/reports/countries/jp.html (last visited Mar. 18, 2022). Finally, a visitor could compare Japan's statistics to any one of the other one hundred and nine countries for which the Center provide reports or with [external] worldwide data. *Id.*

77. *F4a. Data about content restrictions to enforce terms of service*, 2020 RANKING DIGIT. RTS. CORP. ACCOUNTABILITY INDEX, https://rankingdigitalrights.org/index2020/indicators/F4a (last visited Apr. 19, 2022) (reporting that Facebook and Google do not publish data about the total number of pieces of content restricted for violating the company's rules, Twitter provides partial data on this topic, and describing other limitations of current reports); Hannah Bloch-Wehba, *Global Platform Governance: Private Power in the Shadow of the State*, 72 S. METHODIST U. L. REV. 27, 72–74 (2019) (describing limitations in transparency reports including, lack of information about the number of videos Google removed for "violent extremism," lack of clarity about what actions Facebook has taken with respect to various pieces of content, and lack of standardization).

78. *See* Hannah Bloch-Wehba, *Exposing Secret Searches: A First Amendment Right of Access to Electronic Surveillance Orders*, 93 WASH. L. REV. 145, 158–62 (2018) (providing an overview of how the law constrains what the public knows and what platforms can report about government requests for user data); Alex Abdo, *More Transparency Needed For Government's Use of National Security Powers For Data Requests From Companies*, AM. C.L. UNION (June 19, 2012) (explaining that Google's—and other companies—2011 transparency report provided no insight into the use of surveillance authorities, such as National Security Letters, to obtain user data). Companies can now provide information about national security letters and FISA orders in bans of 250. Letter from James M. Cole, U.S. Deputy Attorney Gen., to General Counsels of Facebook, Google, Linkedin, Microsoft, and Yahoo, January 27, 2014 (on file with authors).

79. John Bowers & Jonathan Zittrain, *Answering Impossible Questions: Content Governance in an Age of Disinformation*, 1 HARV. KENNEDY SCH. MISINFO. REV. 1, 5 (2020).

Assessing whether these symbolic structures enhance the legitimacy of the content moderation function—by directly addressing democratic deficits and promoting public values—poses a real challenge. How can we tell whether they are merely symbolic[80]—providing companies with legitimacy without addressing underlying concerns—or examples of a salutary "[p]rocess era of internet governance"[81] that reflects legitimate models for addressing controversial content governance issues that pit individual rights against collective interests?

Simply weighing the negative and positive attributes of individual structures offers little purchase on the question; these questions confound an easy answer at such a high level of generality. The task of content moderation is not an undifferentiated soup in which a dash of "transparency," "participation," or "fairness" can provide the right taste. Assessing whether or these structures "build legitimacy around how content is sorted, filtered, and ranked"[82] requires greater clarity about the standard(s) against which to measure discrete interventions.[83]

A closer look at the foregoing descriptions of the three legality-evoking forms points to a different lens for analyzing the different structures employed by platforms engaged in content moderation. Specifically, rather than assessing these structures as tools that seek to legitimate content moderation function *as a whole*, they in fact represent attempts to address deficits in particular tasks, or subfunctions, that together comprise the broader function.

Understood in this way, Google's Advisory Council was not assigned the power to conduct all aspects of content moderation but to add legitimacy to one subfunction: *defining* rules and policies governing content. The legal framework created liability but gave next to no guidance on how to evaluate the privacy claims of individuals, the competing interests of the public and other stakeholders, or any guidance on how to weigh the two. Absent such guidance, Google was implicitly tasked with the complex and contentious definitional work of moderating content. They attempted to legitimate the definitional and operational work necessary to moderate content by assigning

80. Edelman, *supra* note 8, at 1542.

81. Bowers & Zittrain, *supra* note 77, at 7.

82. *Id.* at 5.

83. Barrie Sander parses out different content moderation activities along the lines of their purpose to support human rights analysis and due diligence. *See* Sander, *supra* note 10, at 998–90 (identifying four platform content moderation activities in which transparency and oversight should be provided: rule making, decision-making, content and advertising, and regulatory compliance).

that subfunction to a diverse group of independent experts deliberating in a transparent and public process.

Facebook assigned to its Oversight Board, in turn, authority over a second content moderation subfunction: the *application* of rules and policies to determine whether specific content meets the platform's governing definition. Such a task involves both construing content in context to determine its meaning and interpreting platform rules and decisional criteria to decide their application. Thus, through the FBOB, Facebook has attempted to demonstrate consistency with some of the core competencies involved in traditional forms of law-application through public adjudication.

Transparency reports, in their current form, provide public transparency into the outcomes of a third subfunction: case *resolution*. Unlike actions taken through court processes, there is limited visibility into the information removed from the web through the processes emerging under today's content moderation regimes. Transparency reports seek to empower a variety of stakeholders by creating a public record of decisions to remove expression from the public view.

Looking at content moderation through this functional lens reflects the reality of the whole range of statutory frameworks that deal with the treatment of content online. While Section 230 provides no explicit guidance regarding the allocation of subfunctions to different actors, numerous other laws reflect an intention by Congress to do so explicitly, in a variety of ways.

Understanding the symbolic forms adopted by platforms through a functional lens suggests new questions to ask when assessing their use in content moderation:

> To what extent (if at all) does Google's enlistment of input from the public into the *definition* of rules and policies governing content provide competencies that can address democratic deficiencies in that subfunction?
>
> To what extent (if at all) does Facebook's assignment of oversight over the *application* of rules in specific situations to experts address concerns about values raised by that subfunction?
>
> To what extent (if at all) does the information about the *resolution* of specific cases provided through transparency reports enlist competencies that remedy deficits raised by that subfunction?

With these questions in mind, the next Part isolates public values attached to the three subfunctions already identified—definition, application, resolution—and adds a fourth: *identification*. It then examines ways that different tasks or subfunctions are allocated to particular actors (private and

public; human and technical) and the competencies that each contribute to achieving those values through the multi-actor system as a whole. Part III will then return to the assessment of the symbolic structures discussed above, using the analytic framework presented below to assess these legalistic forms, and suggest ways that legislative frameworks might better structure corporate content moderation.

III. DEVELOPING A FUNCTIONAL FRAMEWORK

Informed by the understanding that content moderation does not involve a single function, but rather a set of discrete subfunctions, this Part describes four discrete subfunctions. The subfunctions identified are common across diverse content moderation systems, although the choices made about their assignments and structures vary. Specifically, this functional framework identifies the set of discrete subfunctions as:

> (1) the *definition* of the content subject to moderation;
>
> (2) the *identification* of potentially covered content;
>
> (3) the *application* of the definition to identified content; and
>
> (4) the *resolution* of a particular case (including labeling, amplifying, depressing, or removing).

Distinguishing these subfunctions at a granular level permits a more rigorous inquiry into which actor, or combination of actors, might best be tasked with their performance and how the performance of those subfunctions should be structured. This helps pinpoint which governance values are most salient to different stages of the content moderation process; what competencies are required to perform the subfunction consistent with those values; which actor or combination of actors might provide those competencies; and what constraints should be imposed to ensure that those competencies are brought to bear.

This Part uses existing regulatory frameworks to illustrate common subfunctions, variations in their implementation, and ways policy choices shape them. The examples, drawn from Section 230, the Digital Millennium Copyright Act, the General Data Protection Regulation, and the mix of regulations that shape how platforms handle child sexual abuse material, together illuminate the choices policymakers and other stakeholders can make to guide and constrain decision-making using each subfunction. This analysis seeks to identify regulatory frameworks that allocate and constrain content moderation subfunctions to be more or less supportive of the democratic values bound up in the regulation of speech.

A. GROUNDING A FUNCTIONAL APPROACH

Our suggestion that a rigorous assessment of content moderation must focus on the component subfunctions, the different ways those subfunctions are allocated and constrained, and the deficits that arise under them that reflect an absence of governance-related competencies, is grounded in theoretical insights from "New Governance," values-in-design scholarship, and the realities of the current content moderation practices of platforms.

1. Theoretical Roots

A functional approach reflects the insights of our and others' work on "New Governance." [84] This approach recognizes the ways that legal frameworks—from detailed imposition of content standards to liability regimes—delegate, explicitly or implicitly, the content moderation function to private platforms.[85] Accordingly, it focuses on looking both *within* and *across* the black box of networked-organizational decision-making processes responsible for content moderation on a granular level. The goal is to further the alignment of these processes with public governance norms,[86] by taking seriously the choices between distinct human and technical actors within those contexts and the networks they engage. [87] It further draws attention to emerging trans-governmental networks and transnational forms of private regulation, many of which delegate functions of content moderation to technical actors.[88] Extending the focus to private governance activities, and their comportment with public values, can in turn "catalyze the ongoing development of meaningful internal practices."[89]

More specifically, this functional orientation to regulatory allocations in content moderation applies the "handoff model" that one of the authors has

84. *See, e.g.*, Bamberger & Mulligan, *supra* note 20, 480–82 (summarizing the "new governance" literature).

85. *See* Bamberger, *Regulation as Delegation*, *supra* note 20, at 383 (describing how "private firms increasingly exercise regulatory discretion of the type delegated to agencies").

86. *See id.* at 384 (suggesting replacing a "compliance" paradigm for one focused on "accountability"); *id.* at 383 (proposing the need for a richer account of decision-making within the corporate "black box" to understand the extent to which firms' exercise of regulatory discretion is accountable to public norms).

87. *See, e.g.*, Bamberger, *Technologies of Compliance*, *supra* note 20, at 673 (focusing on the use of technology systems in risk management decision-making).

88. *See, e.g.*, Gregory Shaffer, *Theorizing Transnational Legal Ordering*, 12 ANN. REV. L. & SOC. SCI. 231, 239 (2016) (discussing the ways that the delegation of content moderation functions to technical actors, at times placing "public law . . . in the shadow of transnational private regulation").

89. Bamberger & Mulligan, *supra* note 20, at 482.

developed with philosopher Helen Nissenbaum.[90] The handoff model suggests that the values implications of decision-making models cannot be understood by looking generally at an overall "function," like content moderation itself. Rather, it requires isolation, separation, and examination of the subfunctions that comprise the content moderation function and of the ways that such subfunctions are assigned to differently situated humans, technologies, and combinations of the two.

Current debates in content moderation speak to the significance of the ways that platforms enact content moderation through these functional handoffs. For example, particular corporate implementations of the content moderation function have been critiqued for misidentifying content as covered by platform policy due to an inability to account for context that is essential to appropriately applying a definition.[91] Both humans and technical actors have misidentified content, yet the cause of misidentifications stem from different limitations and biases. While scholars have often raised concerns about privatization and automation, the handoff model promotes a more precise exploration of *how* each actor enlisted in a content moderation regime executes the function delegated to it.

Dissecting subfunctions enacted under various content moderation regimes reveals important questions about the quality of performance, efficiency, and the effect on markets. More importantly, here, it directs attention towards the ethical and political questions that arise in alternate versions of content moderation systems. It complicates the external appearances of *sameness*, exposes how different allocations of content moderation functions implicate different values, and permits an analysis of the range of constraints that might accompany specific functions to protect important public values.[92]

---

90. *The Concept of Handoff*, *supra* note 17; *see also* Kenneth W. Abbott & Duncan Snidal, *The Governance Triangle: Regulatory Standards Institutions and the Shadow of the State*, *in* THE POLITICS OF GLOBAL REGULATION 46 (Walter Mattli & Ngaire Woods, eds., 2009).

91. *See, e.g.*, Aarti Shahani, *With 'Napalm Girl,' Facebook Humans (Not Algorithms) Struggle To Be Editor*, NPR: ALL TECH CONSIDERED, (Sept. 10, 2016), https://www.npr.org/sections/alltechconsidered/2016/09/10/493454256/with-napalm-girl-facebook-humans-not-algorithms-struggle-to-be-editor (discussing human reviewers removing historically significant photos that contained nudity under Facebook's "community standards"); Louise Matsakis, *Tumblrs Porn-Detecting AI Has One Job—And It's Bad At It*, WIRED.COM (Dec. 5, 2018), https://www.wired.com/story/tumblr-porn-ai-adult-content/ (algorithms identifying and removing images as "adult content" under Tumblrs policy).

92. As an example, consider the application of a handoff lens to the access control technologies used in various generations of the Apple iPhone. *See The Concept of Handoff*, *supra* note 17, at 242-248. These have progressed from user-selected passwords to Touch ID (the fingerprint recognition system), to Face ID, by which the iPhone camera constructs a 3D map

The insights of regulatory scholars Kenneth Abbot and Duncan Snidal enhance this analysis. They suggest that evaluation of the strengths and weaknesses of multi-actor governance schemes should focus on whether the actors tasked with performing specific tasks possess, or can harness, necessary competencies.[93] Those competencies include independence, representativeness, expertise, and operational capacity, and different sets of them are essential to the legitimacy of distinct governance activities.[94] "It is difficult if not impossible," they write, "for any non-state actor to provide all the competencies on its own. Thus, the most promising strategy may be collaboration: assembling the needed competencies by bringing together actors of different types."[95] Assessing the capacities of the different actors to whom various subfunctions of content moderation are allocated offers a means to assess whether the complete suite of competencies has been brought to bear in a way that harnesses organizations' and specific human and technical actors' competencies and addresses relevant deficits in the emergent, networked, regulatory system.[96]

2. Descriptive Realities

The functional framework further reflects the reality of existing content moderation practices. Expanding the category of "content moderation" to include the vast and diverse array of laws that involve the regulation of online

of a person's face. A typical narrative for framing the substitution of these different mechanisms might emphasize technological progress: the *same* access function is being performed by increasingly sophisticated technological means, resulting in an upward linear trajectory in terms of security and, perhaps, user experience. The handoff lens' emphasis on the systemic relation of both technological and human inputs into the system, by contrast, reveals that the choice of mechanism implicates important differences in terms of values, including human control and agency, transparency, and privacy.

93. Abbott & Duncan, *supra* note 90, at 44–88.

94. *Id.* at 46.

95. *Id.*

96. *See* Robert Gorwa, *The Platform Governance Triangle: Conceptualising The Informal Regulation Of Online Content*, 8 INTERNET POL'Y REV. 1, 13 (2019) (using Abbott and Snidal's "governance triangle" to analyze a range of content governance schemes and noting the need for research focused on the varying regulatory competencies different actors bring to the content moderation).

material—including data protection,[97] civil rights,[98] intellectual property,[99] and other laws, as well as voluntary agreements that drive platforms to moderate content—reveals a host of content moderation practices that operate very differently from the canonical Section 230 framework. While Section 230 provides essentially no guidance regarding the moderation of content, these other regimes feature statutory schemes that are more explicit in allocating different subfunctions to different actors subject to a range of constraints. Focusing on content moderation as a functional output provides a set of in-the-wild case studies regarding the possibilities of more thoughtful assignment of different tasks to achieve different goals, leverage distinct competencies, and mitigate or backfill deficits.

B. THE FUNCTIONAL FRAMEWORK

The handoff model argues that dissecting content moderation into its component parts facilitates an understanding of the concerns about values raised at each step and the ways that different allocations of subfunctions to actors with different constraints can support or undermine values.

---

97. GDPR, 2016 O.J. (L 119) 1, http://ec.europa.eu/justice/data-protection/reform/files/regulation_oj_en.pdf.

98. *See, e.g.*, Title VII of the Civil Rights Act of 1964 §§ 703–16, 42 U.S.C. § 2000e to 2000e-15 (prohibiting discrimination in employment); *id.* at § 2000e-3(b) (prohibiting advertisements that "indicate a preference, limitation, specification or discrimination" with respect to protected classes); Age Discrimination in Employment Act (ADEA), 29 U.S.C. § 623(e) (similar provision prohibiting advertisements that express a preference); Fair Housing Act (FHA), 42 U.S.C. § 3601 (prohibiting housing discrimination against protected classes); *id.* at § 3608 (prohibiting advertisements that "indicate a preference, limitation, specification or discrimination" with respect to protected classes).

99. Transparency reports produced by key platforms, as discussed in Section III.A.3, *infra*, include information about content moderation occurring under a range on intellectual property legal frameworks across jurisdictions, including copyright, trademark, and trade secret. *See, e.g.*, *Copyright Notices*, TWITTER TRANSPARENCY, https://transparency.twitter.com/en/reports/copyright-notices.html#2020-jul-dec (last visited Mar. 18, 2022) (providing separate reports on removals under the DMCA); *Trademark Notices*, TWITTER TRANSPARENCY, https://transparency.twitter.com/en/reports/trademark-notices.html#2020-jul-dec (last visited Mar. 18, 2022) (same under Twitter's trademark policy). Google provides a specific transparency report on requests to delist links from search results based on copyright, and their resolution and their general content removal statistics include content removed "due to claims of trade dress and/or distinctive marks." *Content Delistings due to copyright*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/copyright/overview (last visited Mar. 18, 2022). This includes, but is not limited to, claims of counterfeit and trademark. GOOGLE TRANSPARENCY REP. HELP CTR., https://support.google.com/transparencyreport/answer/7347744?hl=en (last visited Apr. 20, 2022). Facebook's Transparency Center includes a section on actions taken on claims of intellectual property violations, including copyright, trademark, and counterfeit goods. *Intellectual Property*, META, https://transparency.fb.com/data/intellectual-property/ (last visited Mar. 18, 2022).

A given legal scheme may not explicitly designate who is responsible for a particular subfunction. Yet in practice, a different actor or set of actors—public or private, human or technical—performs each subfunction and the allocations that emerge under existing content moderation regimes differ. For example, a social media platform could employ more reactive or proactive measures to initiate cases by relying on users to report potential violations of content policies, training employees to perform the identification task, using an automated system that relies on natural language processing, or a combination of all three.

At times, subfunctions may be encumbered with procedural constraints, such as a requirement to provide notice to a user when removing content, or constraints on the actors who can actually remove content, such as encouraging the use of an automated filter to screen out objectionable content or constraining the use of automation.

Below we define each discrete subfunction, note the core values associated with its performance and the competencies those demand, and use case studies to explore and contrast how different content moderation regimes allocate the subfunctions.

1. *Components of A Functional Framework*

a) Identifying the Subfunctions of Content Moderation

Every content moderation regime explicitly or implicitly assigns responsibility for a set of discrete subfunctions that comprise the content moderation task: definition, identification, application, and resolution.

Definition—Each content moderation regime must set rules or policies defining the type of content it targets. Different regimes afford distinct actors more influence, at least for some period, over the definition of content subject to moderation. The actor who is authorized to craft the definition wields immense power.

While it may be tempting to assume that formal law does the heavy lifting on this important aspect of content moderation policymaking, that is often not the case. A definition may be explicitly captured in statutory language or case law. Yet, even when a statutory definition exists, it may take the form of a multi-factor balancing test or a broad standard that in practice shifts power for defining content to the entity facing liability. Sometimes the decision about which content to moderate is left entirely to platforms, without any requirement that they formally create definitions, leaving the definition of content subject to moderation intuited through removals rather than through *ex ante* definitions or guidelines.

Identification—Once the content to be moderated is defined, some actor(s) must be tasked with initiating inquiries by identifying potentially covered content—the functional equivalent of bringing a legal case. This subfunction might be assigned to the platform itself (through direct commands or liability regimes), to public actors (such as prosecutors, law enforcement agents, and regulatory agencies), or to a range of other private actors (including rights-holders or other parties claiming injury. This range of responsible actors, further, may rely on human efforts to flag relevant content or technical systems using artificial intelligence.

The identification subfunction can also include the subsidiary task of actually locating the content online. In some instances, the process for identifying content includes identifying its location; reporting systems, for example, sometimes require provision of a URL.[100] Yet a content moderation regime might explicitly or implicitly assign the responsibility for identifying and locating content to distinct actors.[101] For example, under the context of the National Center for Missing and Exploited Children database discussed below,[102] once content is identified child sexual abuse material by human analysts, technical system—hash databases of that content—are used to support the location and summary removal of matching images. The ability to bifurcate identifying and locating creates interesting opportunities to leverage the competencies of distinct public and private, and human and technical, actors. Many contentious battles over online content center on whether platforms have a responsibility for identifying and locating regulable content.

Application—Once the content to be moderated is defined and potential instances of covered content is identified, a determination must be made as to whether the content meets the definition. This subfunction involves the application of a rule to a particular fact pattern, a function traditionally assigned to public processes of adjudication, whether judicial or administrative. Such application often involves both interpreting the connotations of the content in context, the meaning of the decisional rules, and the fit between them. Thus, depending on the type of content and relevant rules involved, the legitimacy of such a process requires not just a technocratic application of rules to facts but also constraints that ensure that judgment and interpretive discretion satisfies a range of public values, from democratic oversight, to participation, to consistency, proportionality, and fairness. Achieving this balance has posed

100. *See, e.g.*, *infra* Section II.B.2.b (discussing the right-to-be-forgotten context).

101. In addition, the location of the content may alter the identification. For example, using a video of police brutality that has been identified in a hate group context to locate that video in a news context may alter the outcome of the identification subfunction.

102. *See infra* text accompanying notes 134–139.

an especially thorny task when authority is vested in supra-national, or private, decision-making bodies rather than regulatory or enforcement agencies or public prosecutors.[103]

Different content moderation regimes assign this role to different actors and sometimes to multiple actors over the course of a dispute. Sometimes, as in the case of the Digital Millennium Copyright Act, public courts remain a key actor; in other regimes, such as Section 230, the private sector performs the application task entirely.

Resolution—This subfunction includes decisions about the full range of actions that may be taken once content is identified and located and the decisional rule applied. Resolving cases involves a determination of appropriate remedies. Sometimes the law explicitly or implicitly determines the action, by imposing strict liability for certain content or setting forth detailed statutory provisions directing removal. Other regimes leave platforms free to determine resolution decisions. Common moderating actions include blocking, removal, amplifying, downgrading, flagging, labeling, monetizing, strategically engaging, and reporting (including but not limited to government agencies), and the list continues to expand.[104] While the platform is the actor that most commonly comes to mind as having the right, obligation, or ability to resolve content issues, many platforms provide individual users with the ability to engage in at least some aspects of the task directly, for example enabling users to mute or block content from specific users, to mute specific messages, or to control whether they see content that a platform has designated offensive. Users, moreover, can also construct methods of resolving content issues, creating blocklists and tools that automate them through tools such as Block Bot and the now-defunct Block Together.

The order in which these subfunctions occur can differ by content moderation regime. Frequently, for example, disputed or problematic content is identified, and the rules and policies for determining whether they meet the relevant definition are then applied. But, as in the case of the CSAM moderation regime discussed below,[105] the process of determining that the definition applies to a particular image is sometimes made first, and instances of that image subsequently identified and resolution achieved, without

103. *See*, *e.g.*, Martin Shapiro, *"Deliberative," "Independent" Technocracy v. Democratic Politics: Will the Globe Echo the E.U.?*, 68 L. & CONTEMPORARY PROBLEMS 341 (2005) (discussing the challenge in the transnational context).

104. *See* Eric Goldman, *Content Moderation Remedies*, 28 MICH. TECH. L. REV. 1 (2021), https://ssrn.com/abstract=3810580 or http://dx.doi.org/10.2139/ssrn.3810580 (discussing nearly three dozen moderation actions taken by companies).

105. *See infra* text accompanying notes 130–139.

independent application to the instance of the matched image. Moreover, content moderation regimes might involve the execution of a single subfunction multiple times—often assigning the iterative performances to different actors (whether technical and human, or to different groups of humans)—such as through appeals or review processes that revisit rule-application or case resolution decisions.[106]

b) Examining the Values Implicated, and Competencies Required, by the Different Subfunctions

Deconstructing content moderation through this framework exposes the public values at stake in each, and the different competencies required for legitimate performance. These values and competencies are particularly important because platforms are asked to engage in governance activities that support the "public" or "common" interest[107] as distinct from their business interest.[108] To meet this goal, it is essential to ensure that the allocation of subfunctions align both with the competencies that are critical to procedural aspects of public interest regulation, such as independence and representativeness, and with the expertise and operational capacity necessary for actions and outcomes substantively aligned with the public interest.[109]

Conversations about mixed public-private governance regimes typically focus on questions about the appropriateness of delegating aspects of implementation, or sometimes adjudication, to the private sector. Public law does not usually outsource the core role of policy formation or rulemaking to private entities.[110] Agenda setting and guidance, including establishing

106. *See* Jean Patja Howell, *The Lawfare Podcast: How Zoom Thinks About Content Moderation*, LAWFARE (Dec. 2, 2021), https://www.lawfareblog.com/lawfare-podcast-how-zoom-thinks-about-content-moderation (podcast in which Zoom executives discuss the multiple layers at various stages of content moderation review conducted by different groups).

107. *See* Walter Mattli & Ngaire Woods, *In Whose Benefit? Explaining Regulatory Change in Global Politics*, *in* THE POLITICS OF GLOBAL REGULATION 1, 4 (Walter Mattli & Ngaire Woods eds., 2009) (distinguishing between "captured regulation" which entrenches narrow interests and "common interest regulation" that fulfills broader public purposes). Mattli and Woods argue that public interest regulation can emerge where "open forums, proper due process, multiple access points and oversight mechanisms" exist along with robust societal demand for action. *Id.* at 17.

108. *See* Elkin-Koren & Perel, *supra* note 11, at 858 (distinguishing between "public" functions related to the identification and removal of unlawful content from "private" functions driven by business-related content moderation imperatives).

109. Abbott & Duncan, *supra* note 90, at 3.

110. The Computer Fraud and Abuse Act (CFAA) provides a telling example of the problems that arise where the law makes such a delegation. 18 U.S.C. § 1030. Efforts to bring claims against individuals—researchers, users and employees—who've violated corporate terms of services, which set all sorts of limits on how individuals can interact with systems,

decisional criteria, fall generally within the purview of the public sector with private actors implementing them, exercising only limited discretion. Yet, given the constitutional, jurisdictional, and practical limitations on government efforts to regulate speech that many platforms and many members of society want moderated, platforms currently are doing exactly this core policymaking work. Importantly, even when governments could do it, they have abdicated their role at times, at least in part leaving the core definitional subfunction in private hands.

"Defining" work sets policy and, therefore, particularly around expression, implicates core public values. Substantively, public policy must protect individual rights and balance competing interests. Procedurally, legitimate policy formation requires stakeholder consultation, public participation, expert deliberation, reasoned decision-making, and transparency throughout, including publication of the final policy adopted. Defining work requires competencies: independence (that is, not beholden to one party or having an interest in the outcome), representativeness of relevant stakeholders, and two forms of expertise—substantive (subject matter) and political (to negotiate across stakeholders).

The application of rules and policies implicates different values: the reasonable, consistent, and fair interpretation and application of rules. This work requires particular competencies: independence from parties and from outcome; normative expertise, especially in the global context in which content moderation is enacted; and representativeness. Issues about jurisdiction and juries of peers get at this competence and point to the difficulty of theorizing and operationalizing it in this context.

Constraints on who can perform the "identifying" subfunction, like standing requirements in litigation, limit the category of parties who can raise concerns about content. This may lead to underenforcement of rules, particularly if the remaining parties enabled to do the identifying work are less resourced. Explicit assignment of the "identifying" subfunction can act as a check on the co-option of platforms' technical and human resources by more financially or politically powerful interest groups. In assigning the "identifying" subfunction, regulators should consider the entity's operational capacity and

---

have been limited by court concerns with the substantive and procedural concerns of outsourcing the definition of what is illegal to the private sector. *See* United States v. Drew, 259 F.R.D. 449, 464 (C.D. Cal. 2009) (concluding that construing the CFAA to cover terms of service would render the statute void for vagueness); *see* Van Buren v. United States, 141 S. Ct. 1648, 1662 (2021) (narrowing construction of unauthorized access under the CFAA to cover obtaining information from particular areas within a computer to which an individual does not have access, but not to cover accessing available information for improper purposes).

the extent to which over or under enforcement is more aligned with public values. Certain allocations of the "locating" function may promote interests such as competition, which may be of importance independently or because of the potential to support freedom of expression.

Finally, resolving cases, like applying definitions to particular content, implicates the values of consistency, proportionality, and fairness with regard to penalties and other remedies. There may be more tolerance for variations in how distinct platforms resolve cases, rather than apply governing standards, pursuant to voluntarily adopted policies. Yet concerns with the consistency, proportionality, and fairness of the remedies imposed remain. For these reasons, resolving work, which assigns penalties and other remedies, requires independence from parties and from outcome, as well as normative expertise.

c) Constraints Intended to Protect Values and Enhance Competencies

In some instances, regulators recognize that an entity to whom they are allocating a subfunction is not fully equipped to perform it, or to perform it legitimately. To address foreseen deficits, regulations sometimes direct or constrain how a subfunction is performed. Constraints take two forms: process constraints and limitations on the actors used for implementation.

Process constraints impose procedures on the execution of a subfunction. These range from requirements for transparency or secrecy of rules, policies, or outcomes to detailed rules setting out how the subfunction must be implemented. Actor constraints establish a preference for a specific kind of actor or a limit on what type of party can perform the subfunction—human or technical. The case studies discussed below highlight ways in which such constraints are explicitly imposed as well as places where they are predictable, if not explicitly prescribed, outcome of regulatory design choices.

2. Understanding Elements of the Functional Framework Through Case Studies

Existing content moderation regimes provide a rich set of examples to explore the allocation of subfunctions and the use of constraints. Commentary on different regimes further highlights how particular arrangements of the content moderation function put public values more or less at risk.

a) Case Studies of Subfunctions

i. Defining: Section 230 and the DMCA

Section 230 of Communications Act[111] and the Digital Millennium Copyright Act[112] represent the ends of the spectrum with respect to the allocation of the defining subfunction. While Section 230 does not explicitly define content to be moderated, by shielding platforms from civil liability for removing or blocking content that they consider "obscene, lewd, lascivious, filthy, excessively violent, harassing, or otherwise objectionable," the law gives those private actors wide berth to define the content subject to moderation. The broad protection against liability is neither contingent on the content's illegality—even content that is constitutionally protected may be removed at the platform's pleasure—nor on the platform's motives.

The law also encourages platforms to empower users in moderating content, by limiting civil liability for providing technology that helps users restrict access to objectionable material. As a formal matter, such technology leaves the definition of objectionable to the users themselves; the tools can be used to remove whatever content a user deems objectionable. Yet, in practice, the tools' provision of categories of content for filtering constrain this definitional flexibility. Moreover, such tools, only allow users to configure the content to which they personally are exposed, rather than shape what is in circulation. Together, these two provisions largely give companies the latitude to define through written text, practices, and tools, the content that circulates on their platforms.

In contrast, federal law defines the content to be moderated under the DMCA. While the processes and responsibilities for action set out in the statute give copyright holders the ability to influence what content is removed, the definition of removable content is tethered to the legal definition of copyright infringement.[113] Several provisions are designed to ensure that only infringing material is subject to moderation. There are ex ante checks on copyright holders' claims,[114] and there is recourse to courts to resolve disputes over whether content is infringing. Moreover, while platforms may decide to

111. 47 U.S.C. § 230.

112. 17 U.S.C. § 512.

113. 17 U.S.C. § 512(c)(3) (tethering the use of the takedown process to copyright law, and in particular, requiring the complaining party to have a good faith belief that use is infringing); CoStar Grp., Inc. v. LoopNet, Inc., 373 F.3d 544, 552 (4th Cir. 2004) (indicating that DMCA safe harbor is irrelevant in determining what constitutes a prima facie case of copyright infringement).

114. 17 U.S.C. § 512(c)(3)(A)(v) ("Good faith" attestation); 17 U.S.C. § 512(f) (penalties for misrepresentation); 17 U.S.C. § 512(g) (counter-notice and putback).

reject a takedown request, nothing in the law requires or even invites them to do so. Indeed, the statutory framework protects platforms against liability for removing or limiting access to content based on a copyright holder's complaint or any other good-faith basis, regardless of whether the material is ultimately determined to be infringing or not. Individuals whose content is the subject of a takedown request can object, and if a dispute persists, either party may file a suit in federal court. This recourse to the courts maintains the role of public law and public institutions in defining the content to be moderated under the DMCA.[115] While much of the action under the law occurs in the notice and takedown process outside the courtroom, giving rise to misuse,[116] recourse to the courts, along with other regulatory design choices below, maintain the public role in defining moderated content.[117]

These different statutory allocations of definitional work shape stakeholders' perceptions of the legitimacy of the content moderation activities platforms take underneath them.[118] For example, because § 230 enables

115. The role of the courts proved essential to maintaining the balance under copyright law between the rights holder's interest and the public's interest and, in particular, to establishing that before issuing a takedown notification a complaining party must consider whether the use of the material constitutes fair use because "fair use is 'authorized by the law.' " Lenz v. Universal Music Corp., 815 F.3d 1145, 1153 (9th Cir. 2016) ("We conclude that . . . fair use is 'authorized by the law' and a copyright holder must consider the existence of fair use before sending a takedown notification under § 512(c)."). However, the consideration of fair use requires only good faith, meaning the complaining party can only be held liable if they "knowingly misrepresented" their own understanding of whether the use of the copyright constituted fair use. *Id.* at 1154.

116. Jennifer M. Urban, Brianna L. Schofield & Joe Karaganis, *Takedown in Two Worlds: An Empirical Analysis*, 64 J. COPYRIGHT SOC'Y USA 483, 514 (2017) (concluding based on empirical research that §512 is used to address "privacy, defamation, and other disputes" and to target "non-infringing material"); Jennifer M. Urban and Laura Quilter, *Efficient Process or Chilling Effects? Takedown Notices Under Section 512 of the Digital Millennium Copyright Act*, 22 Santa Clara Computer & High Tech. L.J. 621, 667 (2006) (finding substantive problems with 31% of § 512(c) and (d) notices reviewed); Daniel Seng, *Copyrighting Copywrongs: An Empirical Analysis of Errors with Automated DMCA Takedown Notices*, 37 SANTA CLARA HIGH TECH. L.J. 119 (2020).

117. The statute establishes safe harbors for internet service providers to protect them from liability for copyrighted works others make available through and on their systems. It does not preclude the parties to a dispute about the use of a copyrighted work from accessing the courts. The statute creates an additional remedy under § 512(f) for bad faith issuance of a takedown notice. *See* Lenz v. Universal Music Corp., 572 F. Supp. 2d 1150, 1154–55 (N.D. Cal. 2008).

118. Elizabeth Dwoskin, *Trump is Suspended From Facebook for 2 Years and Can't Return Until 'Risk to Public Safety is Receded,'* WASH. POST (June 4, 2021), https://www.washingtonpost.com/technology/2021/06/03/trump-facebook-oversight-board/ (describing Facebook's response to a ruling from its Oversight Board's ruling on the appropriateness of banning former president Donald Trump as "an attempt to clarify Trump's penalty and make the procedures

platforms to engage in unbounded, undeclared, and unfettered moderation, users find the experience of being moderated mysterious, arbitrary, and at times, biased. In contrast, stakeholders have raised concerns about the copyright holders' desire to ignore the statutory requirement to consider fair use before filing a takedown notice.[119] Moreover, concerns about covered content have not raised questions about the legitimacy of the platforms' actions. Thus, the distinct allocations of the definitional subfunction have contributed to different perceptions of the legitimacy of platforms' moderation activities.

b) Identifying: RTBF and CSAM

The "right to be forgotten" provision in the EU General Data Protection Regulation, which requires search engines to delist search results that violate an individual's privacy interests,[120] and the suite of statutes, including the PROTECT Act,[121] which shape how platforms moderate child sexual abuse material, allocate responsibility for identifying regulable content to different actors. In each regime, the law on the books does not explicitly enlist platforms in the work of identifying content. However, other aspects of the regulatory frameworks incentivize platforms to take on identification work.

The Article 29 Working Party's guidance clarifies that the individual claiming the right to be forgotten bears the responsibility for identifying content for erasing, deleting, or delisting, at least in the first instance.[122] The guidance documents clarify that the law does not require search engines to

---

of the powerful social network, which is used by 3.45 billion people globally on a monthly basis, appear less arbitrary and opaque to the public").

119. *See*, *e.g.*, India McKinney & Ernesto Falcon, *Electronic Frontier Foundation Memo to Incoming Biden Administration*, EFF (Jan. 21, 2021), https://www.eff.org/wp/eff-transition-memo-incoming-biden-administration#_Toc57064038 ("Section 512 strikes a balance between the interests of service providers, copyright owners, and Internet users—but the system is not perfect. It is too easy for copyright owners . . . to have speech taken down, which comes at a high price for free expression and the public interest. The problem of false and abusive takedown notices is widespread and well documented.").

120. GDPR, art. 17, 2016 O.J. (L 119) 1, 43 ("Right to Erasure ('right to be forgotten')") (providing that "[t]he data subject shall have the right to obtain from the controller the erasure of personal data concerning him or her without undue delay and the controller shall have the obligation to erase personal data without undue delay" where one several grounds applies).

121. Pub. L. 108–21, 117 Stat. 650 (2003) (codified at 18 U.S.C. § 2258A).

122. *Guidelines on the Implementation of the Court Of Justice of the European Union Judgment on "Google Spain and inc. v. Agencia Española de Protección de Datos (AEPD) and Mario Costeja Gonzáles" C-131/12*, at 6 (Nov. 26, 2014), http://ec.europa.eu/justice/article-29/documentation/opinion-recommendation/files/2014/wp225_en.pdf [hereinafter *Guidelines Implementing Google Spain Judgment*] ("The ruling does not oblige search engines to permanently carry out that assessment in relation to all the information they process, but only when they have to respond to data subjects' requests for the exercise of their rights.").

proactively identify personal data for removal but only "to respond to data subjects' requests for the exercise of their rights."[123]

However, because the regulation provides no guidance on the format or content of notices, the specificity with which the content being requested for erasure or restriction (delisting) must be identified is uncertain. This uncertainty complicates the interactions between users, hosts, and search engines during the identification subfunction.[124] On one hand, relevant guidance from the Article 29 Working Party directs requesters to "identify the specific URLs."[125] On the other, that same guidance, as well as guidance from member states' data protection authorities, affirms the data subject's right to make erasure requests as they see fit[126] and, in particular, to use methods beyond whatever standardized intake forms entities provide.[127] This combination creates some uncertainty with respect to whether platforms may be required, at the very least, to assist in identifying content covered by the regulation.

Moreover, the law allows an individual to request delisting, or other action by a search engine, without requesting erasure from the host. This complicates the distribution of responsibility for identifying—and the related task of locating—content covered by the law. If a host alters the URL at which content resides, a search engine could end up returning content it has attempted to delist from queries on the name of the data subject. In such situations, it is unclear whether there is any shift in responsibility for identifying or more narrowly re-locating the content subject to delisting. In addition, if the information subject to delisting was publicly shared online by a controller, the controller must take "reasonable" steps to inform other entities of the data

123. *Id.* at 6.

124. The Art 29 Working Group Guidance; guidance issued by the Information Commissioner's Office in the U.K. states that Individuals can make a request for erasure verbally or in writing.

125. *Guidelines Implementing Google Spain Judgment*, *supra* note 122, at 7. The European Data Protection Board issued its own guidelines, *see* Guidelines 5/2019 on the criteria of the Right to be Forgotten in the search engines cases under the GDPR (part 1), (July 7, 2020), https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201905_rtbfsearchengines_afterpublicconsultation_en.pdf, however the Board does not address the content of erasure requests specifically.

126. *Guidelines Implementing Google Spain Judgment*, *supra* note 120, at 7 (noting that "national data protection laws provide for great flexibility . . . and offer data subjects the possibility of lodging their requests in a variety of ways" and so while it may be convenient for data subjects to use forms and procedures online services set up "it should not be the exclusive way for data subjects to exercise their rights").

127. *Id.*

subject's request.[128] This may create a responsibility to assist in locating the content subject to delisting on other platforms.

In contrast, even though the statutory framework governing child sexual abuse material[129] allocates formal responsibility for identification outside the platforms, platforms have taken on some shared responsibility for the identifying subfunction.[130] Some platforms have developed their own databases[131] of "hashes," which are numeric "fingerprints" of previously identified images of child sexual abuse, against which they screen content on upload. Some companies claim to use machine learning classifiers to identify

128. GDPR, art. 17, recital 66, 2016 O.J. (L 119) 1, 13, 43.

129. These include general criminal statutes prohibiting the making, printing, publishing, distribution, reproduction, transportation, and possession of child pornography, 18 U.S.C. §§ 2251, 2251A, 2252, 2252A, 2252B, 2260; as well as the statute establishing the National Center for Missing and Exploited Children (NCMEC), an independent private agency funded by the U.S. Department of Justice's Office of Juvenile Justice and Delinquency Prevention tasked with assisting with the identification and recovery of missing and exploited children. NCMEC coordinates programs to locate missing children, provides technical assistance and training to law enforcement and other stakeholders, and provides information and assistance services. 42 U.S.C. § 5773(b)(1). While existing federal statutes refer to child pornography, we use the term child sexual abuse material which more accurately captures the content of the images which depict the rape, sexual abuse, and sexual exploitation of children and is the preferred term among experts. *See The EARN IT Act: Holding the Tech Industry Accountable in the Fight Against Online Child Sexual Exploitation: Hearing on S. 3398 Before the S. Comm. on the Judiciary*, 116th Cong. 2 n.1 (2020), https://www.judiciary.senate.gov/imo/media/doc/Shehan%20Testimony.pdf [hereinafter Testimony of John Shehan] (testimony of John Shehan, Vice-President, Exploited Children Division, National Center for Missing & Exploited Children).

130. Platforms are not required to proactively search out CSAM for removal, and responsibility for identifying CSAM is left to others: victims, platform users, law enforcement. Where platforms have actual knowledge of CSAM offenses, they are required to file reports with the NCMEC. *See* The Prosecutorial Remedies and Other Tools to end the Exploitation of Children Today Act of 2003 (PROTECT Act), Pub. L. No. 108-21, 117 Stat. 650 (2003) (codified at 18 U.S.C. § 2258A). Penalties for knowing failure to report are $150,000 for the first violation and $300,000 for each subsequent violation. 18 U.S.C. § 2258A(e). The law places mandatory reporting requirements on providers of electronic communication service and remote computing services to the public.

131. *See* Testimony of John Shehan, *supra* note 127, at 3 (stating that "many technology companies . . . actively detect and remove child sexual exploitation content" and "go above and beyond the requirements of current law and look for innovative methods to address child sexual abuse material and implement sophisticated tools and technologies to identify this content online, report it to NCMEC, and get it quickly removed"); United States v. Miller, 982 F.3d 412, 419 (6th Cir. 2020) (describing Google's use of proprietary hashing technology to create hashes of confirmed child sexual abuse images, and scan customer files on upload for matches which a Google employee might view and confirm as child sexual abuse material or might just send an automated report with the file to NCMEC).

CSAM imagery that is not yet cached in hash databases.[132] Machine learning algorithms are trained on known CSAM to identify statistical patterns which are then used to identify potential CSAM in the wild.[133]

Many platforms rely on a shared hash database hosted by the same non-profit entity, the National Center for Missing and Exploited Children (NCMEC),[134] which employs a cross-platform technology called PhotoDNA to scan images. These scans happen generally prior to allowing them to be uploaded, against the NCMEC database of CSAM. According to the inventor of PhotoDNA, Hany Farid, more than 95% of the nearly 18 million reports in 2018 to NCMEC's CyberTipline, constituting over 45 million pieces of identified CSAM, were from photo DNA.[135]

The NCMEC database process further allocates identification tasks between various human and technical actors. Platforms can choose whether or not to share the child sexual abuse material they find on their networks as part of the mandatory reports.[136] Human NCMEC analysts determine whether reported images are CSAM and, if so, add hashes of the images to the NCMEC database.[137] Thus, the company reports provide continuous source material

132. *See, e.g.*, Kristie Canegallo, *Our Efforts to Fight Child Sexual Abuse Online*, GOOGLE BLOG (Feb. 24, 2021), https://blog.google/technology/safety-security/our-efforts-fight-child-sexual-abuse-online/.

133. Hany Farid, *Reining in Online Abuses*, 19 TECH. & INNOVATION 593, 595 (2018).

134. NCMEC is authorized to receive and review the CSAM. 18 U.S.C. § 2258A(a), (b)(4).

135. *Fostering a Healthier Internet to Protect Consumers: Joint Hearing Before the Subcomms. on Commc'ns & Tech. & Consumer Prot. & Commerce, of the House Comm. on Energy & Commerce*, 116th Cong. 2 (Oct. 16, 2019) (testimony of Hany Farid, Professor, University of California, Berkeley).

136. 18 U.S.C. § 2258A(b), (b)(4) (stating that reports "may, at the sole discretion of the provider, include . . . [a]ny visual depiction of apparent child pornography or other content relating to the incident such report is regarding"). From conversations with knowledgeable experts, the authors believe that the major platforms routinely share the images and videos they identify with NCMEC. However, we do not have a citation to support this claim, and we do not know whether other entities behave similarly. While the large platforms engage in active efforts to screen for CSAM, a 2020 NCMEC reported that only 1,400 of the approximately 7,000 electronic service providers who are statutorily required to report had voluntarily registered to report with NCMEC CyberTipline, and of those 1,400, only 169 had actually filed a report in 2019. Testimony of John Shehan, *supra* note 127, at 4.

137. *South Korean National and Hundreds of Others Charged Worldwide in the Takedown of the Largest Darknet Child Pornography Website, Which was Funded by Bitcoin*, U.S. DEPT. OF JUST. (Oct. 16, 2019), https://www.justice.gov/opa/pr/south-korean-national-and-hundreds-others-charged-worldwide-takedown-largest-darknet-child (reporting that previously unknown CSAM is analyzed by NCMEC for potential inclusion in the hash database). As of February 2014, NCMEC analysts only review files attached to a report if the report indicates that either the reporting company reviewed it or that it was publicly available. U.S. DEPT. OF JUST., INFORMATION PAPER FOR PROSECUTORS AND LAW ENFORCEMENT OFFICERS: CYBERTIPS AND SUPPRESSION: AVOIDING AND DEFENDING AGAINST FOURTH AMENDMENT CLAIMS 4

that can be analyzed and used to update the database of *identified* CSAM used by other providers to quickly *locate*, and then remove (the only acceptable form of *moderation* in this context) it from their platform. Similarly, material found on personal hard drives and other media, such as USB flash drives and other portable storage drives, during law enforcement investigations—often triggered by reports to the CyberTipline—is hashed and fed into the NCMEC database.[138]

A complicated set of factors, rather than a regulatory mandate alone, drives the use of technology to identify CSAM. These factors include NCMEC publicizing and encouraging the use of PhotoDNA; Microsoft's decision to donate the technology to NCMEC for free licensing to eligible entities; and pressure from governments, advocacy organizations, and individual victims and families. The statutory framework that requires service providers to report CSAM to NCMEC and eliminates liability for doing so facilitates the creation of the shared database.[139]

Prior to the use of PhotoDNA and other technology by platforms to identify CSAM, the identification task was left to other actors.[140] Given the limitations of the technologies currently in use—hash databases locate

(2021). Prior to February 2014, NCMEC's analysts would "determine if the material constitutes a violation of law." U.S. GOV'T ACCOUNTABILITY OFF., GAO-03-272, COMBATING CHILD PORNOGRAPHY: FEDERAL AGENCIES COORDINATE LAW ENFORCEMENT EFFORTS, BUT AN OPPORTUNITY EXISTS FOR FURTHER ENHANCEMENT 9 (2002). This shift responded to a decision holding NCMEC to be a state actor where it opened and examined a file attached to a CyberTipline Report that the provider had not reviewed. *See* U.S. v. Keith, 980 F. Supp. 2d 33, 41–42 (D. Mass. 2013). Reviewing material that a provider has previously reviewed falls within the private search exception, whereas reviewing material that has not been reviewed by a human at the reporting entity may not. *See* U.S. v. Reddick, 900 F.3d 636 (5th Cir. 2018) (holding that identifying photos through a PhotoDNA match was a private search and that law enforcement human review of those files did not exceed the private search doctrine). The question of whether NCMEC is a governmental entity or state actor shapes the current review process. In *U.S. v. Ackerman*, 831 F.3d 1292, 1306–1307 (10th Cir. 2016), the court held that NCMEC was a government entity or a state actor and its review of the emails and attachments contained in a CyberTipline Report that not been examined by the provider violated the Fourth Amendment. Subsequent legislation has attempted to clarify that NCMEC is a private non-profit entity, but it is unclear whether it has done so. *See* CyberTipline Modernization Act of 2018, Pub. L. No. 115-395, 132 Stat. 5287 (Dec. 21, 2018).

138. Interview with Hany Farid, Professor, U.C. Berkeley Sch. Info, on file with authors, June 26, 2020.

139. 18 U.S.C. § 2258A (establishing reporting requirements and penalties for failure to report); *id.* § 2258B (limiting liability for required reporting).

140. Given that platforms are not required to screen content for violations of law, they generally rely on users—and others—to identify (and generally require them to simultaneously locate) content that violates both law and platform policies. They use a range of techniques to do this, ranging from reporting forms to flags.

previously identified CSAM, and predictive classifiers are imperfect—the torrent of CSAM, and the rise of end-to-end encrypted services, the public (victims, advocacy organizations, etc.) and law enforcement continue to play an important role in identifying CSAM.

The different allocations of the identifying subfunction under the RTBF and the CSAM frameworks have produced various critiques. While the CSAM regime implicitly allocates the task of identifying content to be moderated on other stakeholders, as described above, it does not require platforms to establish procedures to assist individuals in reporting CSAM. The reporting infrastructures for CSAM on large platforms vary and have been criticized for being difficult to find and use, for failing to support victims,[141] and for failing to remove CSAM consistently. Reporting structures for CSAM are more difficult to find and use than those for copyright, for example.[142] Platforms do not have specific processes or flags within general content reporting processes to report alleged CSAM material. Within major platforms, "in nearly all cases it was impossible to explicitly flag content as CSAM."[143] The complexity and inconsistency between desktop and mobile versions of the same platforms across different platforms make identifying CSAM material for action difficult. In many instances, the reporting functions are insensitive to the nature of the crime and victims—for example requiring reports to come through platform accounts, requiring identifying information, and requiring information about the alleged perpetrator without clear guidance about how it will be used. Similarly, the lack of clarity about who and how content subject to action under the right to be forgotten should be identified skews incentives. In the CSAM context, the lack of standardized identification requirements and process, particularly when contrasted with the clear process set out under the DMCA, burdens victims who are relatively less powerful, have less capacity, and are at risk to harm that cannot be remedied through damage awards. The allocation of identification work without clear procedures, in both the CSAM and RTBF context, has created hurdles for victims and their allies, as well as platforms, and contributed to concerns about the substantive commitment of platforms to addressing the substantive harms at issue.

141. CANADIAN CTR. FOR CHILD PROT., REVIEWING CHILD SEXUAL ABUSE MATERIAL REPORTING FUNCTIONS ON POPULAR PLATFORMS 7 (2020) (reporting that survivors of CSAM generally characterize the reporting experience as "disheartening," lengthy, and sometimes futile, and they report being challenged by moderators).

142. *Id.* ("[U]sers concerned with issues related to copyright infringement, [sic] almost universally have access to formal reporting tools and clear instructions for initiating a complaint," none of which are available for CSAM victims.)

143. *Id.* (describing research findings with respect to Twitter, Facebook, Microsoft, Google, Snapchat, TikTok, Discord, Pornhub, and Xvideos).

The NCMEC hash database described above[144] presents an example of the way that assignment for identifying content as problematic, and actually locating it online, can be bifurcated. While the statutory framework does not specifically allocate responsibility for either identifying or locating CSAM, it facilitated the emergence of a shared identification infrastructure in two ways. First, the statutory framework creates a centralized collection of images by requiring electronic service providers to report alleged CSAM to the NCMEC CyberTipline and eliminating any risk of liability for sharing CSAM material.[145] Second, the law allows NCMEC to share hash values of collected CSAM (but not the images themselves) with other electronic service providers for the exclusive purpose of stopping the sexual exploitation of children.[146]

After material is identified and found to meet the definition (*application* subfunction discussed below) of CSAM, a hash of the image is included in the database. Participating entities receive the hashes of all CSAM content in the NCMEC database—those they have contributed, and those others have contributed—to aid in locating matching content on their networks. In this framework, material that is a candidate for moderation is sent to NCMEC whose analysts determine whether it is CSAM (rule *application*), and if so, add it to the hash database.[147] The hashes are then used to locate that same image or video (or a slightly perturbed version of it) on many platforms to facilitate further removal, reporting, and subsequent investigations.

Discretizing the subfunctions of identifying and locating has several potential benefits. First, NCMEC staff and law enforcement, as described above, apply the rules by determining what material enters the database.[148] These entities may be perceived as more expert, and therefore their decisions about whether rules apply to specific material are more legitimate. Second, platforms, other electronic service providers, and scholars have noted that proactive screening for CSAM material is costly and reviewing such images is emotionally difficult for employees.[149] The shared database of known CSAM

144. *See supra* text accompanying notes 134-138.

145. 18 U.S.C. § 2258(B).

146. 18 U.S.C. § 2258(c)(a). NCMEC also shares reported information with law enforcement. 18 U.S.C. § 2258(C)(d) (allowing NCMEC to make reports, including images, available to law enforcement).

147. The NCMEC database contains a subset of CSAM which its creators decided was indisputably illegal: images of children under the age of 12, who are typically prepubescent, involved in an explicit sexual act. Hany Farid, *Reining in Online Abuses*, 19 TECH. & INNOVATION 593, 598 (2018).

148. *Supra* text accompanying notes 134-138.

149. U.S. GOV'T ACCOUNTABILITY OFF., GAO-03272, COMBATING CHILD PORNOGRAPHY: FEDERAL AGENCIES COORDINATE LAW ENFORCEMENT EFFORTS, BUT AN OPPORTUNITY EXISTS FOR FURTHER ENHANCEMENT 2002 [hereinafter COMBATING CHILD

enables all platforms to benefit from the prior identification and reporting work of peer platforms and NCMEC's application work. The automated screening tool reduces the costs and improves the pace of content moderation and reduces the human toll of identification work. Third, the task of identifying and locating can impose a disproportionate burden on smaller companies. Sharing the benefits of previous identification and application work, and easing the burden of locating through automation, can reduce the cost and labor associated with moderating CSAM.[150] This shared infrastructure for CSAM can help platforms behave in ways that other stakeholders view as legitimate and socially responsible.

The allocations of content moderation subfunctions in CSAM has raised substantive and procedural concerns around privacy and due process. The NCMEC database allows platforms to allocate some identification work to NCMEC while sharing the location work through the PhotoDNA software, which allows them to screen content against it. Yet this intertwining of functions, combined with questions about whether NCMEC is a government agent or actor, has raised constitutional concerns.[151] Courts in the United States have reached different opinions about whether or not this structure creates a special relationship between NCMEC and law enforcement.[152] Congress has attempted to eliminate these concerns, and NCMEC has altered the way it handles images to avoid constitutional privacy concerns more clearly.

c) Applying: Section 230 and the DMCA

Section 230 and the DMCA provide examples of different ways the task of applying the rules to content identified as potentially subject to moderation can be allocated. Section 230 sits at one end of the spectrum, the DMCA at

---

PORNOGRAPHY]; George W. Burruss, Thomas J. Holt, & April Wall-Parker, *The Hazards of Investigating Internet Crimes Against Children: Digital Evidence Handlers' Experiences with Vicarious Trauma and Coping Behaviors*, 43 AM. J. CRIM. JUST. 433 (2018); Kathryn C. Seigfried-Spellar, *Assessing the Psychological Well-Being and Coping Mechanisms of Law Enforcement Investigators vs. Digital Forensic Examiners of Child Pornography Investigations*, 33 J. POLICE & CRIM. PSYCH. 215 (2018).

150. DEPT. FOR DIGIT., CULTURE, MEDIA & SPORT & HOME OFF., ONLINE HARMS WHITE PAPER 56 (2019) ("Badly designed regulation can stifle innovation by giving an advantage to large companies that can handle compliance more easily. We are determined that this regulatory framework should provide strong protection for our citizens while avoiding placing an impossible burden on smaller companies."); COMBATING CHILD PORNOGRAPHY, *supra* note 147 (reporting on industry feedback reporting that cost is a barrier to developing and using technology to proactively identify CSAM).

151. Tyler O'Connell, *Two Models of The Fourth Amendment and Hashing to Investigate Child Sexual Abuse Material*, 53 U. PAC. L. REV. 293, 309–14 (2021).

152. *Id.*

the other. As described above, the law does not obligate companies to adopt or enforce any rules about content, and to the extent platforms do so, they are free to apply them however they like, leading many to view the statute as an invitation to abdicate good governance.[153] However, this view provides only a partial picture. While the law places no affirmative obligations to moderate content, the drafters' goal from inception was to remove legal obstacles to platform content moderation.[154]

By shielding platforms from liability for activities that would create liability as a publisher or secondary publisher (distributor) in the offline world, the law empowers platforms to engage in establishing rules and applying them to remove, edit, or otherwise moderate content without incurring any liability. Section 230's restraint on civil liability protects behind the scenes work that limit what content is available online and the development of settings and tools which users can use to moderate content. In both instances, the platform is determining the content to which its rules apply.

On the other end of the spectrum sits the DMCA. Unlike Section 230, which lumps entities who engage in a wide range of distinct functions into the single category of interactive computer service providers, the DMCA delineates different types of service activity in relation to third party content—hosting, locating, caching, transmitting—and specifies whether and what sort of actions entities must take to avail themselves of the safe harbor protections. For example, the statute directs requests for the removal of content to platforms that host content rather than search engines or caching services.[155]

More importantly, while the law requires one of these functionally defined entities (hosts) to serve as a messenger between parties contesting whether the

153. *See, e.g.*, Dawn C. Nunziato, *The Death of the Public Forum in Cyberspace*, 20 BERKELEY TECH. L.J. 115, 118 (2005) (arguing that "the government's abdication of control over Internet speech regulation may well result in the loss of protection for speech that is insufficiently protected within an unregulated market for speech"); Danielle Keats Citron & Helen Norton, *Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age,* 91 B.U. L. REV. 1435, 1456 (2011) (noting that § 230 allows companies to decide whether and how to shape online expression and that while many opt to govern online hate speech, many others have not and some have built businesses around tolerating or encouraging online hate speech).

154. *See* JEFF KOSSEFF, THE TWENTY-SIX WORDS THAT CREATED THE INTERNET 63–64 (2019);*see also id.* at 57–76 (discussing drafters' goals more generally).

155. DMCA, 17 U.S.C. § 512(b)(2)(E) (requiring caches to remove material only if it has previously been removed from the originating site or access to it has been disabled, or a court has ordered that the material be removed from the originating site or that access to the material on the originating site be disabled; and the party giving the notification includes in the notification a statement confirming that the material has been removed from the originating site or access to it has been disabled or that a court has ordered that the material be removed from the originating site or that access to the material on the originating site be disabled).

use of a copyrighted work is infringing, it leaves the application of copyright rules to that content to the parties or, if the parties choose, the courts. During the dispute between the parties, moreover, the statute explicitly controls how social media platforms, and others who host user generated content, should treat content—directing them to take it down expeditiously upon the receipt of a notice, restore it if a counter notice is received and the copyright holder does not notify the platform that they are seeking a court order for removal, and specifying timelines for the later actions.[156] The law shields social media platforms from liability for removing or limiting access to content based on a copyright holder's complaint or any other good faith basis regardless of whether the material is ultimately determined to be infringing as long as they follow the safe harbor rules.[157] Both parties to a dispute have recourse to federal courts to resolve the claimed infringement, maintaining, to some extent, the role of public law in the application work of content moderation. Critiques of allocations of the application function provide useful insight into how constraints can address concerns with substantive and procedural legitimacy. Advocates and scholars have criticized platform moderation protected by § 230 for being inconsistent and opaque. Procedural critiques include the lack of transparency about rules and their application and the lack of an appeals process. Criticism has also focused on the actors platforms use to apply rules. Human actors tasked with moderating content may over- and under- block because as Sarah Roberts details in her research, both the leads and the front-line workers are often geographically and culturally removed from the platform they are monitoring.[158] Given this geographic and cultural removal, platform workers may lack the tacit knowledge necessary to fairly apply definitions to specific content.[159] Technical tools (in handoff parlance actors) are poor at accounting for information about culture, context, or use

156. *Id.*

157. 17 U.S.C. § 512(g)(1).

158. SARAH T. ROBERTS, BEHIND THE SCREEN 35 (2019) ("Both the headquarter from the tech platform and its audience may be very far removed, geographically and culturally, from the location where workers are viewing and moderating the user-generated content.").

159. Given the lack of visibility into content moderation practices it is difficult to assess the extent to which various factors, including geographically and culturally diversity, affect content moderation decisions. One recent qualitative study documents sensitivity to cultural differences relevant to content moderation, difficulties in reconciling, and training efforts. *See* SABRINA AHMAD, "IT'S JUST THE JOB": INVESTIGATING THE INFLUENCE OF CULTURE IN INDIA'S COMMERCIAL CONTENT MODERATION INDUSTRY 21 (2019) (finding Indian content moderators engage in a "holistic and critical *analysis* of this content for accuracy and legitimacy").

that in many cases are essential to determining whether an item of content meets the definition of content to be moderated.[160]

d) Resolving: § 230 and the DMCA

As with rule application, Section 230 and the DMCA also illustrate two ends of the spectrum with respect to case resolution. Under the DMCA, the only resolution actions are removal and restoration, and the statutory framework tightly controls the timing and other requirements of both. There is no discretion left to platforms and therefore a limited critique of those resolutions. In practice, however, some platforms offer copyright holders other options for resolving disputes. For example, when a video on YouTube is surfaced through the Content ID program, which features a database of copyrighted works submitted by owners, copyright holders may choose monetization over removal. When videos are uploaded to YouTube, they are scanned against these files, and when a match occurs, the video may be blocked or the owner may instead choose to run advertisements before the video plays and reap the revenue. Through Content ID, copyright holders can further leave the video on YouTube but restrict which applications or websites can embed it. The safe harbor provisions explicitly limit the moderation activities of entities that allow users to transmit and locate information and those who cache information—for example entities that allow users to transmit information may not modify the content or choose recipients. The resolution activities of hosts are also shaped by statutory provisions that mirror general contributory and vicarious liability standards.[161]

Section 230, in contrast, leaves platforms free to determine the range of appropriate resolutions. They are protected from liability for any resolutions—

160. Emma Llansó and her co-authors similarly distinguish between the use and risks posed by (1) artificial intelligence of various sorts used for proactive detection (what we call identification) of content potentially subject to moderation on the one hand and (2) automated evaluation and enforcement (what we call application and resolution) of the policies governing identified content on the other. They identify risks across both subfunctions such as false positives and false negatives; biased and/or discriminatory performance; escalating need for private data; inadequate human involvement and oversight; as well as specific risks of automated enforcement (application and resolution) including normalizing prior restraints on speech, the lack of due process, limits on human oversight, and limits on redress and accountability. *See* Llansó, *supra* note 3, at 9–10.

161. 17 U.S.C. §§ 512(c)(1)(A), (B). Abiding by the provision of the safe harbors has protected platforms from liability for contributory and vicarious liability. However, specific kinds of advertising and customer support has supported successful claims against platforms under inducement. *See* Metro-Goldwyn-Mayer Studios Inc. v. Grokster, Ltd., 545 U.S. 913, 936–37 (2005) (finding liability for inducing infringement where a platform marketed itself as a venue for sharing copyrighted material, explicitly encouraged users to upload copyrighted materials, and actively assisted them in doing so).

removing or otherwise limiting access to content they deem objectionable—taken in good faith, as well as any resolutions that information content providers or users make through technical tools platforms provide.[162] As Jeff Kosseff explains in his book on the history of Section 230, "…companies will not be considered to be the speakers or publishers of third-party content, and they will not lose that protection only because they delete objectionable posts or otherwise exercise good-faith efforts to moderate user content."[163] Courts have held that the "otherwise objectionable" and "good faith" language largely protect platforms' discretion.[164]

C. LESSONS FROM THE CASE STUDIES: THE TYPES OF CONSTRAINTS USED IN STRUCTURING SUBFUNCTIONS

These legal frameworks also illustrate the types of constraints used to structure the performance of different subfunctions in ways that promote relevant values and bring relevant competencies to bear. These include constraints on the process used in decision-making and constraints on the allocation of functions between particular technical and human actors.

a) Process Constraints

Sometimes statutory frameworks anticipate the ways in which allocations of subfunctions may put the rights and interests of other parties, or the public interest, at risk. They accordingly establish some combination of procedural

162. 47 U.S.C. § 230(c)(2) (shielding interactive computer services and users from liability for removing or limiting access to content or providing technical tools to help information content providers or others to restrict access to content); *see, e.g.*, E360insight, LLC v. Comcast Corp., 546 F. Supp. 2d 605, 607–08 (N.D. Ill. 2008) (protecting Comcast against Internet marketing company's tort claims arising from Comcast's use of filters to block unsolicited emails).

163. KOSSEFF, *supra* note 154, at 65–66.

164. Barnes v. Yahoo!, Inc., 570 F.3d 1096, 1105 (9th Cir. 2009) (barring a claim against Yahoo! for failing to remove nude photos after promising to do so); Domen v. Vimeo, Inc., 433 F. Supp. 3d 592, 603–04 (S.D.N.Y. 2020) (stating that § 230(c)(2) "does not require that the material actually be objectionable; rather, it affords protection for blocking material 'that the provider or user considers to be' objectionable"); Enigma Software Grp. USA, LLC v. Malwarebytes, Inc., 946 F.3d 1040, 1051–52 (9th Cir. 2019) (concluding that "if a provider's basis for objecting to and seeking to block materials is because those materials benefit a competitor" their action would fall outside (c)(2) but that "the catchall ["otherwise objectionable"] was more likely intended to encapsulate forms of unwanted online content that Congress could not identify in the 1990s" and therefore provided interactive computer services with the ability to moderate content deemed objectionable beyond the "seven specific categories that precede" the term.) For an in-depth discussion of § 230(c)(2) case law, see Edward Lee, *Moderating Content Moderation: A Framework for Nonpartisanship in Online Governance*, 70 AM. U. L. REV. 913, 971–81 (2020).

constraints or limitations on the use or kind of automation or human actors used to implement the subfunction. The DMCA, for example, sets procedural constraints on many subfunctions of content moderation.[165] The law establishes relatively rigorous and balanced procedural rules that constrain the identification, application, and resolution subfunctions. The statute dictates the information that copyright holders must provide to initiate the takedown process (identification). It requires platforms and other hosts to designate an agent[166] to receive notices of alleged infringement. It also requires platforms to be transparent about the actions they take under the law, including notifying the party who posted the content that it has been removed, sharing the takedown notice that triggered removal, and informing them of their right to file a counter notice challenging the allegations in the takedown. It controls the timing of various aspects of the process, requiring platforms to "expeditiously" remove content when they receive such a notice[167] and to reinstate the content in "not less than 10, nor more than 14" days if the user files a counternotification and the complaining party does not notify the platform that they are filing a court action in response.[168] While the platform must expeditiously take down the content alleged to be an infringement, they must also promptly notify the party who posted the content to maintain protection against claims flowing from that removal. The subfunctions delegated to various parties are constrained through procedural protections aimed at protecting the interests of all parties to the dispute and, in particular, ensure that the opposing parties have access to basic information necessary to go to court. The procedural requirements thus ease the burdens placed on platforms, ensure that copyright holders have access to predictable and standardized complaint processes across platforms, and assure that users

165. *See* text accompanying at *supra* notes 112–118.

166. 17 U.S.C. § 512(c)(2) (stating the person's contact information must be available on the website and from the Copyright Office).

167. 17 U.S.C. § 512(c) (stating they must meet a set of other criteria to be eligible for the safe harbor including they do "not have actual knowledge" of a copyright infringement).

168. 17 U.S.C. § 512(g)(2)(b).

receive information in a timely manner so they can assert their rights and are protected from frivolous,[169] overbroad,[170] and false claims.[171]

Procedural constraints come in many forms. They may be designed to ease the work placed on platforms or to protect the rights of competing individuals by regularizing processes and information flows, or they may tilt the field in ways that favor the interests of specific parties.

b) Constraints on the Allocation of Functions Between Particular Technical and Human Actors

Some content governance regimes require or prohibit certain kinds of actors from undertaking specific subfunctions. This may be affected through a requirement or incentive to use a particular protocol or database or a prohibition on automating certain subfunctions, at least when the results have particular effects on individuals.

---

169. 17 U.S.C. § 512(c)(3). They must be in written form and signed by "a person authorized to act on behalf of the owner" of the copyright; and contain a statement that the information is accurate, and under penalty of perjury, that the complaining party is authorized to act on behalf of the owner of an exclusive right that is allegedly infringed.

170. 17 U.S.C. § 512(c)(3)(A)(v) (providing that a valid DMCA complaint must include a "statement that the complaining party has a good faith belief that use of the material in the manner complained of is not authorized by the copyright owner, its agent, or the law."). The Ninth Circuit applies a subjective standard for "good faith," meaning "a copyright owner cannot be liable simply because an unknowing mistake is made," even if the mistake was unreasonable. Rossi v. Motion Picture Ass'n of America, Inc., 391 F.3d 1000, 1005 (9th Cir. 2004). Instead, "there must be a demonstration of some actual knowledge of misrepresentation on the part of the copyright owner." *Id.* The Ninth Circuit has also held that the complaining party must consider whether the use of the material constitutes fair use before issuing a takedown notification. Lenz v. Universal Music Corp., 815 F.3d 1145, 1153 (9th Cir. 2016) ("We conclude that . . . fair use is 'authorized by the law' and a copyright holder must consider the existence of fair use before sending a takedown notification under § 512(c)."). However, the consideration of fair use falls under the umbrella of good faith, so the Ninth Circuit also applies a subjective standard; the complaining party can only be held liable if they "knowingly misrepresented" their own understanding of whether the use of the copyright constituted fair use. *Id.* at 1154.

171. 17 U.S.C. § 512(f). In the case of a knowingly mistaken complaint, the DMCA creates a legal cause of action for the recipient of a false complaint to obtain a remedy from the issuer of the complaint. Because of the subjective standard that courts apply, a plaintiff bringing an action under § 512(f) must demonstrate that the complaining party knew they were issuing a false takedown notice. The Ninth Circuit allows the plaintiff to show "willful blindness" as a means of demonstrating subjective misrepresentation. *Lenz*, 815 F.3d at 1155. To prove willful blindness, a plaintiff must satisfy two factors: first, that the defendant subjectively believes "that there is a high probability that a fact exists," and second, "the defendant must take deliberate actions to avoid learning of that fact." *Id.* (citing Global-Tech Appliances, Inc. v. SEB S.A., 563 U.S. 754 (2011)). In a § 512 (f) action, the plaintiff must show that the defendant knew there was a high probability that the content's use was authorized and that the defendant deliberately avoided learning about its authorization. *Id.*

Rather than simply delegating responsibility to an entity, such constraints prescribe the allocation of workflows in an assigned subfunction between technical and human actors. For example, although § 230 does not limit or direct covered entities to use technology or rely on human judgment in specific ways, Congress intended to spur the market-driven development of filtering tools along with self-regulatory policies to address objectionable content.[172] The law's stated objectives include "encouraging the development of technologies which maximize user control over what information is received by individuals, families, and schools" and "remov[ing] disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material."[173] While it was drafted in part to spur the development, deployment, and adoption of "user empowerment" and "blocking and filtering" technologies[174] and to limit the burdens of human review that attach to publishing in traditional media, it does not explicitly require or constrain automation. To the extent that platforms provide technology to content creators to help them restrict access to material, the law limits their civil liability for providing it. These provisions, like the limitations on liability for third-party speech, were designed to incentivize companies to support technical approaches to content moderation.[175]

Together, these provisions of Section 230 sought to spur the development of shared and proprietary content moderation infrastructures. At the time of its passage there was immense effort to develop technical standards,[176] rating systems, stand-alone products, and built-in "parental controls" to support

172. 141 CONG. REC. H8460 (daily ed. Aug. 4, 1995) (statement of Rep. Cox). For background on the purpose and history of the law, see Brief of Law Professors with expertise in Internet Law, as Amici Curiae Supporting Reversal, Stephen J. Barrett M.D. v. Ilena Rosenthal ¶ 4–11, 51 Cal. Rptr 3d 55 (2006) (No. S122953).

173. 47 U.S.C. § 230(b)(3)–(4).

174. *Id.* Purposes of the act include "to encourage the development of technologies which maximize user control over what information is received by individuals, families, and schools who use the Internet and other interactive computer services" and "to remove disincentives for the development and utilization of blocking and filtering technologies that empower parents to restrict their children's access to objectionable or inappropriate online material . . . ." *Id.*

175. 47 U.S.C. § 230(b)(3) (establishing that "the development of technologies which maximize user control over what information is received by individuals, families, and schools" as a policy goal).

176. *Platform for Internet Content Selection*, W3C, https://www.w3.org/PICS/ (last visited Mar. 24, 2022). The World Wide Web Consortium was developing the Platform for Internet Content Selection (PICS) which provided a specification to enable labels (metadata) to be associated with online content that rating services and filtering software could use. For a description of PICS, see ACLU v. Reno, 929 F. Supp. 824, 838–39 (E.D. Pa. 1996).

content moderation by end users. Some of these tools relied on content creators to identify their work as falling within a definition, while others relied on third parties to identify and label content, and some supported both.[177] The general approach regardless was to put decisions about what to moderate in the hands of end users. This goal was imperfectly met, as many tools provided very limited information to end users about how they defined and identified content.[178]

Thus, Congress intended to provide individual users discretion over content to be moderated, and to allow companies to use and offer content moderation tools, to the extent the law allowed companies to shift power for defining, identifying, and moderating content to individuals, especially parents. Yet in practice the legislation shifted these tasks to third party tool providers. As with platform moderation policies and practices, these third party tools provided customers or users, as well as other stakeholders, with limited information about definitional criteria and implementation details, and were routinely found to engage in over blocking.[179] Content creators were often unaware that their material was being moderated, and if they objected, had no clear path to contest it.[180] While the drafters of the law understood that technology would play an important role in content moderation, they surely did not foresee the vast roles technology plays today or the unique opportunities and challenges it creates.

The DMCA also speaks to the use of technological actors. First, it requires platforms to "accommodate[s]" and "not interfere" with certain "standard technical measures" used to identify or protect copyrighted works.[181] While it

177. For a description of the various filtering and rating technologies available at the time see Am. C.L. Union v. Reno, 929 F. Supp. 824, 839–42 (E.D. Pa. 1996), *aff'd*, 521 U.S. 844 (1997).

178. *See Fahrenheit 451.2: Is Cyberspace Burning?*, ACLU (Aug. 1997), https://www.aclu.org/other/fahrenheit-4512-cyberspace-burning?redirect=fahrenheit-4512-cyberspace-burning (providing a critique of the filtering technologies available at the time); Marjie Nouwen, Nassim Jafarinaimi, & Bieke Zaman, *Parental Controls: Reimagining Technologies for Parent-Child Interaction*, 15th PROC. EUROPEAN CONF. ON COMPUTER-SUPPORTED COOPERATIVE WORK—EXPLORATORY PAPERS, REP. OF THE EUR. SOC'Y FOR SOCIALLY EMBEDDED TECHS. 1, 2–3 (2017), https://dl.eusset.eu/handle/20.500.12015/2928 (describing the four key functions of contemporary parental control technologies: time restrictions, content restrictions, activity restrictions, and monitoring and tracking).

179. For a collection of early studies and tests of Internet filters identifying over-blocking, see generally MARJORIE HEINS & CHRISTINA CHO, INTERNET FILTERS: A PUBLIC POLICY REPORT in FREE EXPRESSION POLICY PROJECT, NATIONAL COALITION AGAINST CENSORSHIP (2001).

180. *Id.*

181. To qualify for safe harbors an entity must "accommodate[] and . . . not interfere with standard technical measures." "[S]tandard technical measures' means technical measures that

does not require copyright holders to use technical components to identify or protect their works, if they choose to, this non-interference requirement means hosting platforms must accommodate them on their infrastructure. Today, copyright holders routinely rely on technology to identify copyrighted works.[182] Second, the law does not condition eligibility for the safe harbors on "monitoring. . .or affirmatively seeking facts indicating infringing activity,"[183] and case law has affirmed that platforms do not need to use technology to monitor or identify infringing material.[184] Thus while the law places a non-interference requirement on entities that host content, it does not require those same entities to independently invest in or use technology to monitor interaction with copyrighted material.

Finally, the GDPR as a whole contains a limit on completely automated decision-making. This provision could limit the capacity of platforms to use technical actors, or at least constrain how they are used, for specific subfunctions. Even where a decision is not based on a fully automated process or meets an exception to the prohibition on solely automated decision-making, entities must conduct an impact assessment of automated decision-making systems that pose a "high risk" to an individual's rights and freedoms prior to adoption.[185] They must also provide individuals with explanations of the decisions such systems render. Guidance from the European Data Protection Board (EDPB)—an independent EU advisory body composed of representatives of the EU national data protection authorities and the European Data Protection Supervisor[186]—on the GDPR principle of "privacy by design" emphasizes the need for evaluations of bias in algorithms that

---

are used by copyright owners to identify or protect copyrighted works and—have been developed pursuant to a broad consensus of copyright owners and service providers in an open, fair, voluntary, multi-industry standards process; are available to any person on reasonable and nondiscriminatory terms; and do not impose substantial costs on service providers or substantial burdens on their systems or networks." 17 U.S.C. § 512(i).

182. Jennifer M. Urban, Joe Karaganis, & Brianna L. Schofield, *Notice and Takedown: Online Service Provider and Rightsholder Accounts of Everyday Practice*, 64 J. COPYRIGHT SOC'Y. USA 371, 374 (2017) (finding that automated "bots" are routinely used by large rights holders to search out infringements and generate takedown notices).

183. 17 U.S.C. § 512(m)(1).

184. *See* EMI Christian Music Grp., Inc. v. MP3tunes, LLC, 844 F.3d 79, 91 (2d Cir. 2016) (holding safe harbor protection could not be conditioned on service provider monitoring its service or affirmatively seeking facts indicating infringing activity); Viacom Int'l, Inc. v. YouTube, Inc., 676 F.3d 19, 34 (2d Cir. 2012) (holding safe harbor protection could not be conditioned on affirmative monitoring by service provider).

185. *See* GDPR, art. 35, 2016 O.J. (L 119) 1, 53–54.

186. EDPB was created by the GDPR to replace the Article 29 Working Party. *See Article 29 Working Party*, EUR. DATA PROT. BD., https://edpb.europa.eu/about-edpb/more-about-edpb/article-29-working-party_en (last visited Mar. 18, 2022) [hereinafter *Article 29 Working Party*].

automate decision-making and the need for human oversight.[187] EDPB guidelines on targeting on social media impose specific requirements on the use of automated systems in certain contexts, affecting how technical actors are used in moderation.[188] Scholars have discussed the importance of these provisions in the context of automated content moderation.[189] More specifically, the provision of the Article 29 Working Party Guidance on the *Google Spain* decision which prohibits platforms from requiring individuals to funnel requests through particular forms or processes can be viewed as another constraint on the automation of the moderation process. Together, these provisions will influence the use of technology to enact some of the subfunctions of content moderation, although how and to what extent remains open.

Applying a functional framework to assess content moderation structures, then, involves asking three questions. (1) What subfunction is the structure intended to bolster or perform? (2) What are the public governance values implicated by that subfunction, and what is the set of competencies necessary to perform it? (3) What constraints, regarding the actor(s) used by the entity to which the subfunction is assigned, and the process they must use best protect relevant values and direct necessary competencies to the task? The next Part will consider the three symbolic legal structures discussed in Part II in light of

187. *Guidelines 4/2019 on Article 25 Data Protection by Design*, ¶ 70 (Apr. 2019), https://edpb.europa.eu/sites/default/files/files/file1/edpb_guidelines_201904_dataprotection_by_design_and_by_default_v2.0_en.pdf (explaining that privacy by design requires qualified human intervention capable of uncovering biases in automation in accordance with Article 22; that algorithms must be regularly assessed to assure they are fit for purpose and to identify and mitigate biases; and, that data subjects should be informed about the functioning of the processing of personal data based on algorithms that analyze or make predictions about them).

188. *Guidelines 8/2020 on the targeting of social media users*, ¶¶ 85, 86 (Aug. 2020), https://edpb.europa.eu/system/files/2021-04/edpb_guidelines_082020_on_the_targeting_of_social_media_users_en.pdf (stating that while prior guidance has found that "targeted advertising based on profiling will not have a similarly significant effect on individuals. . . . However, it is possible that it may do, depending upon the particular characteristics of the case, including: the intrusiveness of the profiling process, including the tracking of individuals across different websites, devices and services; the expectations and wishes of the individuals concerned; the way the advert is delivered; or using knowledge of the vulnerabilities of the data subjects targeted"); *see also Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679*, at 22 (Aug. 22, 2018), https://ec.europa.eu/newsroom/article29/items/612053/en.

189. Llansó, *supra* note 3, at 13 (arguing that human oversight over automated decision-making in the context of content moderation is important at both the "individual decision level and through review of the systems that produced the error" because it "can provide a crucial safety net for the rights and freedoms of affected users").

these questions and suggest ways that they might be used to structure content moderation going forwards.

IV. APPLYING THE FUNCTIONAL FRAMEWORK

The functional framework teases out (1) the values at stake in different subfunctions of content moderation; (2) the competencies required to ensure those values are protected when private actors govern speech in the online public sphere; and (3) the allocations and constraints aimed to address those competencies. This deeper understanding of the values at stake in privatizing different subfunctions provides a tool for assessing existing content moderation structures and constructing new ones.

In this vein, this Part applies the functional framework. It first assesses whether, or to what extent, the legally inspired structures adopted by platforms might either be empty symbols or meaningful efforts to address democratic subfunction deficits. Second, it illustrates how regulatory choices informed by the detailed understanding of the various ways subfunctions can be both allocated to leverage existing competencies and constrained to address deficits, and it illustrates how these choices can contribute to forward-looking decisions about content moderation processes that protect public values.

A. EVALUATING THE SYMBOLIC STRUCTURES

As described in Part II, platforms have fashioned a set of very visible structures in an attempt to backfill the democratic deficits created by particular allocations of subfunctions under current content moderation regimes. Google, through its Advisory Council, convened a body of diverse, independent experts to provide them with advice about *definitional* work based on their expertise, discussions with additional experts around the world, and public feedback. This diversified the expertise and expanded the demographics of the community shaping the important definitional work they were required to perform.

Facebook, through its oversight board, allowed greater scrutiny of the implementation of its rules in concrete cases—the *application* subfunction. By subjecting a subset of its content moderation activities to an independent review process it sought to associate its processes with the values of consistency, proportionality, rationality, and impartiality integral to the adjudicatory process.

Transparency reports have addressed the invisibility of some content moderation *resolution*. Creating a record of content removals and the legal bases behind them provides individuals and the public with an increased understanding of the policies, parties, and politics shaping the information

accessible online. Reports often provide detailed breakdowns on the kinds of removal requests,[190] on the countries from which removals are requested, and some information on the compliance with requests.[191]

The functional framework provides a starting point for more rigorous assessment of whether these symbolic structures, which surely nod towards these important public values, substantively address the democratic deficits associated with content moderation. This allows for analysis of the extent to which they actually build the requisite competencies to attend to the values at play in the discrete subfunctions.

190. *See, e.g.*, *Content Removal Requests Report*, MICROSOFT CORP. SOC. RESP., https://www.microsoft.com/en-us/corporate-responsibility/content-removal-requests-report?activetab=pivot_1:primaryr3 (last visited Apr. 19, 2022) (stating that Microsoft breaks down its content removal into copyright, "right to be forgotten," government requests (covered in their Content Removal Report), child sexual exploitation and abuse imagery, terrorist and violent extremist content, and non-consensual intimate imagery (covered in their Digital Safety Report); GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/ (last visited Apr. 19, 2022) (showing that Google provides separate reports on de-listings under copyright, government requests, European privacy law, YouTube's community guidelines, and the Network Enforcement Law); TWITTER TRANSPARENCY, https://transparency.twitter.com/ (last visited Apr. 19, 2022) (showing that Twitter breaks content removals under their TOS out along many dimensions including violence, terrorism and violent extremism, child sexual exploitation, abuse and harassment, hateful conduct, and provides a separate report about removal requests which covers court orders, information identified by trusted reporters as illegal under local law, among others); *Transparency Reports*, META, https://transparency.fb.com/data/ (last visited Apr. 19, 2022) (Facebook breaks down its reporting into three general categories, actions taken under their "community standards," requests related to intellectual property, and "content restrictions" which includes removal requests from "governments and courts, as well from non-government entities such as members of the Facebook community and NGOs").

191. Facebook, Google, Microsoft and Twitter each provide some information about requests within specific countries and compliance with them, but the detail provided varies. Facebook provides per country aggregates and some illustrative case studies. *See Restrictions by county*, META, https://transparency.fb.com/data/content-restrictions/country (last visited Mar. 24, 2022). Twitter provides per country reports. *See Japan*, TWITTER TRANSPARENCY, https://transparency.twitter.com/en/reports/countries/jp.html (last visited Mar. 24, 2022). For analysis, see *Removal Requests*, TWITTER TRANSPARENCY, https://transparency.twitter.com/en/reports/removal-requests.html#2020-jul-dec (last visited Mar. 24, 2022). Microsoft provides country-by-country breakdowns of requests. *See* MICROSOFT CORP. SOC. RESP., *supra* note 189. Google provides country-by-country breakdowns as well as notable examples. *See Government requests to remove content*, GOOGLE TRANSPARENCY REP., https://transparencyreport.google.com/government-removals/overview?removal_requests=group_by:requestors;period:&lu=removal_requests (last visited Mar. 24, 2022).

1. Google Advisory Council

The *Google Spain* decision and now the General Data Protection Regulation's Right to be Forgotten provisions leave platforms, including Google, with very little guidance on how to avoid liability. This regulatory configuration has the effect of implicitly allocating core definitional work—formulating rules about what content is subject to moderation—along with implementation responsibility (the application subfunction) to platforms.

Critics of Google's delisting system allege that Google serves as a private adjudicator, serving as "the judge and the jury." This is, of course, an outcome of the regulatory framework, which leaves both process and substance largely to Google's judgment.[192] However, this critique doesn't capture concerns over the more unique and consequential delegation to private power: the policymaking, or definition-setting, function.

As Part II describes, the content moderation regimes emerging under the GDPR and the Digital Millennium Copyright Act sit at opposite ends of the spectrum when it comes to definitional work. The DMCA requires platforms to assist in implementation but the task of defining remains within the purview of public law. Platforms are not tasked with evaluating requests beyond their compliance with statutorily prescribed formalities or interpreting the law. In contrast, the GDPR gives platforms the initial responsibility for evaluating requests under a vague and ambiguous law.[193] While complaining parties can ultimately appeal to court if they dislike the platform's ruling, those whose content has been removed lack similar recourse.[194] This gives private platforms a much more central role in the content-governance process. As Edward Lee argues, in the right to be forgotten context, Google acts as a "private administrative agency exercising quasi-lawmaking, quasi-adjudicative, and quasi-enforcement powers."[195] Thus Google is both designing decision-making criteria under a broad legislative standard—akin to the detailed policy

192. Eldar Haber, *Privatization of the Judiciary*, 40 SEATTLE U. L. REV. 115, 137 (2016).

193. *See* Edward Lee, *Recognizing Rights in Real Time: The Role of Google in the EU Right to Be Forgotten*, 49 U.C. DAVIS L. REV. 1017, 1035 (2015) (asking "what institution should have the primary responsibility of addressing or clearing up those ambiguities?" and concluding that "Google has played a defining role in operationalizing the right to be forgotten and deciding what circumstances warrant a removal of a link to personal information or not" and that Google is delegated much authority).

194. Daphne Keller, *The Right Tools: Europe's Intermediary Liability Laws and the EU 2016 General Data Protection Regulation*, 33 BERKELEY TECH. L. J. 287, 359 (2018) (discussing GDPR art. 57(1)(f) which requires Data Protection Authorities to review claims based on data protection rights and concluding that a claimant cannot bring a free expression right challenge to a Right to be Forgotten action before them).

195. *Id.* at 1066.

work done by administrative agencies—and establishing the processes through which those rules are applied.[196] As Jean-Marie Chenou and Roxana Radu write, "the 'right to be forgotten' transforms a private intermediary into a quasi-legal decision-making body in charge of interpreting and implementing a law . . . ."[197]

Google's reliance on external experts and solicitation of public input in crafting policy is responsive to the tasks they were asked to undertake.[198] In forming the council and seeking advice through a broad consultative process, Google was backfilling to address the lack of clear procedural and substantive rules to guide their decisions.

Yet the implementation was fraught from the start. Some viewed the creation of the Advisory Council itself as an effort to undermine public deliberation and guidance. In particular, it was viewed by some as an effort to usurp the implementation guidance role played by the Article 29 Working Party,[199] which did eventually provide guidance in November 2014, a few months prior to the final report of the Advisory Committee. For example, Paul Nemitz, Director of Fundamental Rights in the European Commission, stated that Google could have "quickly, without fuss" implemented the ruling if they merely "outsourced [the takedown requests] to a normal midsize law firm in every member state" but chose instead to "start all this circus" which he viewed as "a very smart PR exercise."[200] Julia Powles argued that Google established the Advisory Council "to provide recommendations in parallel with, and in competition to, democratically legitimate regulators."[201]

Relatedly, Powles and others noted that Google's charge to the Advisory Council and subsequent implementation constrained the deliberations of the

196. Christopher Kuner, *The Court of Justice of EU's Judgment on the "Right to be Forgotten": An International Perspective*, EJIL:TALK! (May 20, 2014), https://www.ejiltalk.org/the-court-of-justice-of-eus-judgment-on-the-right-to-be-forgotten-an-international-perspective/ (stating that "[t]he judgment requires data controllers . . . to strike a 'fair balance' between these rights . . . but gives almost no criteria for doing so").

197. Chenou & Radu, *supra* note 34, at 96.

198. *See* Lee, *supra* note 193, at 1071–72 (describing Advisory Council process and report as akin to a public agency's rulemaking or recommendations and as a quasi-legislative function).

199. The Article 29 Working Party was an independent E.U. advisory body composed of all Member State data protection agencies. It was replaced, under the GDPR, with the EDPB, composed of representatives of the E.U. national data protection authorities and the European Data Protection Supervisor (EDPS). *See Article 29 Working Party*, *supra* note 186.

200. *See* Simon Davies, *Google's "Right to be Forgotten" Offensive Goes Spectacularly Off the Rails*, THE PRIVACY SURGEON BLOG, (Sept. 27, 2014), http://www.privacysurgeon.org/blog/incision/googles-right-to-be-forgotten-offensive-goes-spectacularly-off-the-rails/ (reporting public comments by Paul Nemitz),

201. Julia Powles, *The Case That Won't Be Forgotten*, 47 LOY. U. CHI. L.J. 583, 591 (2015).

Advisory Council from the outset.[202] This context places the Advisory Council in a different light: rather than an effort to fill in work the government allocated to them, it suggests that the Advisory Council was in fact a move to usurp, or at least destabilize, some of the responsibility for defining work implicitly *allocated to a public authority*.

Regardless of the political posture, even in the best light some argued that the Advisory Council fell short on several dimensions essential to its legitimacy.

First, some claimed that the composition of the council lacked the robust stakeholder representation required of government advisory bodies. In particular, some argued that the data protection perspective—and in particular the perspective of the EU data protection authorities—was not well represented. Isabelle Falque-Pierrotin, the head of Article 29 Working Party, argued that the process constituted strategic theater, stating that Google "want[s] to be seen as being open and virtuous, but they handpicked the members of the council, will control who is in the audience, and what comes out of the meetings."[203] While the inclusion on the Council of José-Luis Piñar, the former Director of the Spanish Data Protection Agency and former Vice-Chairman of the European Group of Data Protection Commissioners, challenges this narrative to some extent, scholars reviewing the creation and activities of the Advisory Council found that the process sidelined public authorities. More generally, the Advisory Council members were viewed as having a speech-protective orientation, with few members of the Council supporting the *Google Spain* decision.[204] Finally, some participants noted that the news coverage, which at times misconstrued the ruling, furthered the biases of the proceedings.[205]

Second, the Advisory Council had limited insight into the challenges faced by Google and other stakeholders. The charter of an advisory committee

202. *Id.* at 595–99 (discussing how Google shaped understanding of the ruling and the guidance provided by the Advisory Council).

203. Leila Abboud, *Google Hosts Meetings Across Europe on Privacy Rights*, REUTERS, (Sept. 8, 2014), https://www.reuters.com/article/ctech-us-google-privacy-idCAKBN0H308I20140908.

204. *See* Chenou & Radu, *supra* note 34 at 90–91 (describing perception that advisory council was dominated by individuals with a vested interest in a narrow construction of the right to be forgotten and, although not compensated for participation, other financial and policy entanglements with Google).

205. *See* Oral Testimony of Dr. Evan Harris, Trustee of Article 19, ADVISORY COMMITTEE PUBLIC MEETING, London, England (Oct. 16, 2014) ("[A]ny academic study of press articles would find probably a ratio of 9:1 opposition to the ruling, and to the rights that are supposedly due to be protected in that ruling. And that's the right of the publishing world to do that, they have a vested interest.").

supporting government activity generally establishes that it will receive the necessary support—both logistical and informational—to fulfill its mandate. Here, however, Google provided the Advisory Council with little information about its internal practices, leading a group of influential academics to write an open letter to Google about the process, demanding the release of data to inform the conversation.[206] Pointing out the challenges of expert deliberation without access to critical information, Julia Powles wrote, "the expert hearings are largely conducted in a vacuum . . . leaving the debate to inapt analogies, rather than tales of real, human concern that inspire proactive responses."[207] Absent context, she concluded, the expert body insulated Google's "processes with a veneer of authenticity and respectability" yet lacked real influence over them.[208]

Finally, the Advisory Council's engagement with stakeholders and the public has been critiqued as relatively thin and performative.[209] On the one hand, the creation of the Advisory Council and the seven public consultations in European capitals is a rather atypical effort at publicizing and formalizing efforts to bring in outside perspectives to inform firm policy. While platforms consult experts—including academics and civil society organizations—with some regularity, they typically do so behind closed doors and pursuant to non-disclosure agreements. The publicness of this consultation process, and its breadth, enabled a different level of stakeholder engagement. Numerous journalists, academics, and civil society thought leaders publicly commented on the proceedings, debates, and reports.[210]

On the other hand, Julia Powles persuasively argued:

> [T]hese hearings and their constrained format—where eight speakers were each given a short ten-minute window to present their high-level, often rather vehement, views—in practice, served as a

206. *See* Ellen P. Goodman, *Open Letter to Google From 80 Internet Scholars: Release RTBF Compliance Data*, ELLEN P. GOODMAN (May 14, 2015), https://ellgood.medium.com/open-letter-to-google-from-80-internet-scholars-release-rtbf-compliance-data-cbfc6d59f1bd.

207. Julia Powles, *Google's Grand European Tour Seeks To Map Out The Future Of Data Ethics*, THE GUARDIAN (Sept. 10, 2014), https://www.theguardian.com/technology/2014/sep/10/google-europe-explain-right-forgotten-eric-schmidt-article-29.

208. Julia Powles, *How Google Determined Our Right to be Forgotten*, THE GUARDIAN (Feb. 18, 2015), https://www.theguardian.com/technology/2015/feb/18/the-right-be-forgotten-google-search.

209. *See* Chenou & Radu, *supra* note 34, at 92 (concluding that the Advisory Council "failed to engage public authorities and a wide range of viewpoints").

210. For a compilation of academic commentary on the *Google Spain* decision, *supra* note 26, as well as the Advisory Council, see *Academic Commentary: Google Spain—Compiled by Julia Powles and Rebekah Larsen*, CAMBRIDGE CODE, http://www.cambridge-code.org/googlespain.html (last visited Mar. 24, 2022).

> vehicle for individuals and organizations to express their discontent at various aspects of the ruling, as well as fermenting animosity to the ruling in press coverage.[211]

In short, the public engagement with experts did not facilitate a balanced exploration of views but rather generated more heat than light, and the public engagement with experts cast that heat in one direction. Finally, as described in Part II, although public comment was solicited, there is little evidence that it shaped the Advisory Council's deliberations. While additional experts and the public were provided some opportunity to speak, the process and final report provide little evidence that such participation and input meaningfully informed the deliberations.

2. Facebook Oversight Board

Mark Zuckerberg teased the public with the possibility of a board, made its creation a public spectacle, and has fashioned its activities to maximize its symbolism. Building an independent oversight board to review whether Facebook has accurately, fairly, and consistently applied its rules to particular content responds directly to missing competencies associated with the application subfunction—bringing in diverse, independent experts, creating transparent reviews, and building a body of precedent which creates predictability for stakeholders over time, thus regularizing and publicizing the application of rules. Yet, at every turn, the Facebook Oversight Board (FBOB) has been subject to criticism.

Some of that criticism manifests at the meta-level. Critics argue that Facebook chose a symbolic structure that focused attention on individual decisions—the application subfunction—to distract the public from the more important questions about the overall governing policies—the subfunction of *defining* the applicable rules. For example, the *Washington Post* Editorial Board called for Facebook to grant the Oversight Board greater ability to play an "advisory role" in initial content removal, rather than only coming in when a removal is appealed, since Facebook has faced the most controversy for the content it has refused to take down.[212] Members of Congress pushed members

211. Powles, *supra* note 201, at 594.

212. Editorial Board, *Will Facebook's oversight board actually hold the company accountable?*, WASH. POST (May 17, 2020), https://www.washingtonpost.com/opinions/will-facebooks-oversight-board-actually-hold-the-company-accountable/2020/05/17/e1d46f50-93cd-11ea-9f5e-56d8239bf9ad_story.html ("For the time being, the board can only rule on material that has been removed from Facebook, not material that has remained on the site despite protestations [unless Facebook makes a referral]. Yet it is exactly these 'leave-ups' that catch the company the most flak."). Kara Swisher went further arguing that "solving the problem of how to deal with speech across the largest and most unwieldy communications platform in

of the FBOB to demand authority to provide policy guidance to Facebook "to address the systemic amplification of divisive, racist, and conspiratorial content."[213] They also asked the FBOB to obtain the power to publicly report metrics on their progress.[214] In addition, some felt this orientation drove attention away from legal reforms necessary to establish speech and privacy protection rules.[215] In addition, the timing of the FBOB's creation was viewed as strategically designed to signal commitment yet forestall the adoption of important time sensitive, substantive protections.[216] Progressive advocacy organizations claimed that Facebook intentionally delayed the creation of the Board so as to avoid any board involvement in content decisions around the 2020 election.[217]

Evaluated on its own terms as an effort to provide independent oversight and review of the *application* of Facebook's rules to a limited set of cases subfunction, the FBOB built competencies that make it more than merely symbolic. First, several provisions protect the independence of the FBOB members, and while the initial set of members were hand-selected by Facebook,[218] future members will be selected without Facebook input. The

---

human history . . . may be beyond the capabilities of anyone," noting that asking the oversight board to handle challenging content situations on a case-by-case basis "is trying to push back the ocean with one hand." Kara Swisher, *Who's Up for the Job of Decontaminating Facebook?*, N.Y. TIMES (May 6, 2020), https://www.nytimes.com/2020/05/06/opinion/facebook-independent-oversight-board.html.

213. Letter from the H. Comm. on Energy & Com. to Catalina Botero-Marino, Dean, Universidad de los Andes (Aug. 11, 2020), https://energycommerce.house.gov/sites/democrats.energycommerce.house.gov/files/documents/Botero-Marino.2020.8.11.%20Letter%20re%20Facebook%20Oversight%20Board.CAT_.pdf [hereinafter Letter to Catalina Botero-Marino] ("[W]e worry that your presence on the Board may help legitimize an entity that likely will have no ability to stop Facebook from amplifying conspiratorial and divisive content in search of advertising revenues.").

214. *Id.*

215. Editorial Board, *supra* note 211 ("[T]he responsibility for addressing [privacy and content concerns] shouldn't lie with the board, and it shouldn't even lie with Facebook: Governments ought to set rules of their own, and [the United States'] legislature has fallen short.").

216. *Advocacy Groups Release Open Letter Urging Facebook Oversight Board Members To Take A Stand*, ACCOUNTABLE TECH (July 20, 2020), https://accountabletech.org/media/press-release/ (complaining that the FBOB will not be operational until after the high-stakes 2020 elections).

217. *Id.* (reporting on the letter which also called on the five American members of the oversight board to step down).

218. *See* Klonick, *supra* note 53, at 2456–67 (discussing selection process, including close observation of potential members during table top simulations during consulting phase of FBOB development to determine whether they met criteria FB felt important, including open-minded, active listening, issue matter expertise, and dealing with adverse opinions); *Oversight Board Charter*, *supra* note 59, at art. 1, § 8 (describing initial set of 11 members who would then

independence of the initial set is protected by provisions that (1) allow removal only by the Trust established by Facebook to fund and oversee the Board; (2) allow removals only for a violation of policy rather than for the substance of decisions; and (3) provide for up to three, three-year terms.[219] The Trust was provided with only enough capital to support its operation through two three-year terms,[220] raising questions of the Board as a whole's independence—would funding be renewed if Facebook does not approve of decisions?—and its long term viability. Members of Congress expressed concerns that the Board would not be sufficiently empowered to make independent, meaningful decisions[221] and that it might act as a "smokescreen," allowing Facebook's executives to continue to make final decisions regarding content removal.[222] Members of Congress urged further assurances of board members' independence, asking them to commit to resigning if Facebook failed to grant them meaningful powers.[223] Kate Klonick, who was provided special access to the process leading up to the FBOB creation, believes that the structure will allow the FBOB "to develop and maintain intellectual independence in the long-term."[224]

While the charter constrains the decisions the FBOB can review, within its jurisdiction, many features protect the values at risk in the privatization of adjudication. The FBOB has the capacity to control its docket, an important aspect of independence, follows processes that provide opportunities for parties to be heard, and makes many of its actions transparent to the public.

---

select others over time to fill out the Board); Klonick, *supra* note 53, at 2461 (discussing divergence from charter, and ongoing involvement of Facebook in selection process).

219. *Oversight Board Charter*, *supra* note 59, at art. 1, § 8.

220. Elizabeth Culliford, Facebook pledges $130 million to content oversight board, delays naming members, REUTERS (Dec. 12, 2019), *available at* (https://www.reuters.com/article/us-facebook-oversight/facebook-pledges-130-million-to-content-oversight-board-delays-naming-members-idUSKBN1YG1ZG) (describing Facebook's allocation of an irrevocable grant of $130 million).

221. Maggie Miller, *Facebook Oversight Board to Address Racist, Voter Suppression Content*, THE HILL (Aug. 11, 2020, 07:10 PM EDT), https://thehill.com/policy/technology/511591-house-democrats-pressure-facebook-oversight-board-to-address-racist-voter; Letter to Catalina Botero-Marino, *supra* note 213.

222. Letter to Catalina Botero-Marino, *supra* note 213, at 2 ("[W]e worry that your presence on the Board may help legitimize an entity that likely will have no ability to stop Facebook from amplifying conspiratorial and divisive content in search of advertising revenues.").

223. *Id.* at 4.

224. Klonick, *supra* note 55 at 2484 (concluding that Facebook's exclusion from the process, along with the disqualification of current and former employees, and the procedural affordances for the FBOB and Trust will assure its independence).

The Board has the discretion to accept and refuse reviews from all parties.[225] The FBOB makes a subset of content removals more transparent to the public. Unlike the dry statistics provided by the transparency reports, the FBOB decisions provide a view into the nitty gritty of content moderation work done by Facebook itself.[226] The FBOB is directed to publish and make decisions publicly accessible in a database and report statistics on the number and type of cases reviewed, cases submitted by different regions, and the timeliness of their reviews.[227] While it was unclear under the bylaws whether requests from Facebook would be similarly published, and publicly archived, that is the FBOB practice. While covering a limited set of removals, the transparency is more useful in assessing how and why Facebook removes content. The FBOB decisions are binding on the particular piece of content at issue and establish precedent to guide future decisions.[228] Facebook is required to implement decisions, unless doing so "could violate the law."[229]

The limits on the FBOB's jurisdiction are significant, however, and undermine the substantive value of the FBOB. While the FBOB is charged with "review[ing] content and issu[ing] reasoned, public decisions,"[230] it "cannot review removals under local law,[231] and the FBOB reviews a minuscule number of the millions of instances in which Facebook applies rules to content.[232] The language of the charter would allow users to challenge both

225. *Oversight Board Charter*, *supra* note 59, at art. 2, § 1.

226. Facebook Oversight Board's decisions provide detailed facts about actions Facebook has taken, including content removals, under corporate polices, and determines whether its actions accord with policy. *Board Decisions*, OVERSIGHT BOARD, https://oversightboard.com/decision/ (last visited Mar. 24, 2022).

227. OVERSIGHT BOARD BYLAWS, art. 2, § 2.3.2, https://www.oversightboard.com/sr/governance/bylaws/ (last visited Mar. 28, 2022).

228. *Oversight Board Charter*, *supra* note 59, at art. 2, § 2

229. *Id.* at art. 4.

230. *Id.* at art. 5 § 1.

231. *Id.* at art. 7.

232. In the third quarter of 2021 alone Facebook reports removing 13.6 million pieces of content for violating the violence and incitement policy and 9.2 million pieces for violating rules against bullying and harassment content.*Q3 2021 Community Standards Enforcement Report*, META (Nov. 9, 2021), https://transparency.fb.com/data/community-standards-enforcement/?from=https%3A%2F%2Ftransparency.facebook.com%2Fcommunity-standards-enforcement. In the second quarter of 2021 Facebook reported removing 20 million pieces of content from Facebook and Instagram globally for violating their policies on COVID-19-related misinformation. Guy Rosen, *Community Standards Enforcement Report, Second Quarter 2021*, META (Aug. 18, 2021), https://about.fb.com/news/2021/08/community-standards-enforcement-report-q2-2021/.

removal and non-removal decisions,[233] but the bylaws narrow its jurisdiction by limiting reviews to single pieces of organic content that have been removed from Facebook and Instagram.[234] The FBOB can only review a removal that has exhausted Facebook's internal review process.[235] Finally, individuals must "have an active Facebook or Instagram account" to appeal to the Board.[236] Facebook can ask the FBOB to review "questions related to the treatment of content beyond whether the content should be allowed or removed completely,"[237] and the charter provides Facebook access to an expedited review process in exceptional circumstances.[238]

Together, these limitations create barriers to FBOB review of many consequential removals and non-removals and limit who can seek redress. While all courts have jurisdictional limits and standing requirements, the rationales for the limits set in the FBOB charter and bylaws—and in particular the narrowing between the two—have not been articulated or subject to discussion and debate. There is public concern with what Jack Balkin calls *collateral censorship*[239]—regulations that use private organizations to regulate the speech of others—and the ongoing tension between company commitments to international human rights norms and responding to local law. In this light, limiting FBOB's review of all actions taken under local law sidelines them in areas of profound importance, as the Board must direct its energy toward Facebook's private rules rather than engagement with public governance.

Finally, the FBOB leaves stakeholders in the position of largely taking whatever policies and processes the company offers. Consider the following:

The charter states that "[t]he purpose of the board is to protect free expression by making principled, independent decisions about important pieces of content and by issuing policy advisory opinions on Facebook's content policies."[240] However, the emphasis on reviewing decisions to remove specific pieces of content limits the extent and manner to which questions of how those policies conform to freedom of expression, human rights norms

233. *See Oversight Board Charter*, *supra* note 59, at art. 2, § 1 ("[A] request for review can be submitted to the Board by either the original poster of the content or a person who previously submitted the content to Facebook for review"); Klonick, *supra* note 55, at 2463.

234. OVERSIGHT BOARD BYLAWS, *supra* note 227, at art. 3, § 1.

235. *Oversight Board Charter*, *supra* note 59, at art. 2, § 1.

236. OVERSIGHT BOARD BYLAWS, *supra* note 227, at § 1.2.2. Somewhat bizarrely, access to the Board is explicitly excluded for content and account holders in other Facebook services, including WhatsApp, Messenger, Instagram Direct, and Oculus. *Id.* § 1.2.1.

237. *Oversight Board Charter*, *supra* note 59, at art. 2, § 1.

238. *Id.* at art. 3, § 7.2.

239. Jack M. Balkin, *Old-School/New-School Speech Regulation*, 127 HARV. L. REV. 2296, 2298–99 (2014).

240. *Oversight Board Charter*, *supra* note 59, at 2 (Introduction).

generally, and other public values can arise. Like transparency reports, a key function of the FBOB is to channel public attention in particular ways—to focus on individual outcomes rather than fundamental questions about the appropriateness of Facebook's policies and approaches to implementing them at scale. While the FBOB moves beyond the symbolic, providing protection for core values in the extremely limited and narrow set of cases it reviews, it channels public attention away from an arguably more important set of questions about the legitimate creation of rules—*defining* subfunction—and the entanglement of private platforms in government action. While independent review of platforms application of rules bolsters the protection for freedom of expression and other human rights, the limits on the FBOB's jurisdictional scope, and the fraction of removals it can actually review,[241] limit its substantive value.

3. Transparency Reports

In their current form, transparency reports build operational capacity allowing companies to provide the public with improved understanding of the content removals that limit the information available on the web. However, they are partial and strategic, illuminating the impacts of a slice of the subfunction of *resolution*. They often provide information on a limited set of content removals—specifically the fact that content was removed pursuant to a particular law in a particular country—and provide limited data on that set. They do not reveal the identities of the actor requesting removal or the reasoning behind a removal decision. Thus, while they provide information about general trends and practices at a company over time, and in this way produce data similar to that of the wiretap reports they emulate, they fall far short of the records produced by court actions.

Some scholars view transparency reports as a "major step"[242] in providing "horizontal transparency"[243]—allowing stakeholders a glimpse into the corporate black-box.[244] Yet even fans of the transparency reports note their "major limitations."[245] The quantitative emphasis of the reports leave stakeholders swimming in data but unable to discern its meaning. For example, in 2018 Facebook began including data about removals by their

241. *See* Klonick, *supra* note 55, at 2490 (reporting that based on 2019 figures that "approximately 170,000 pieces of content per day that would be potentially eligible for Board review").

242. Gorwa & Ash, *supra* note 6, at 298.

243. Hans Krause Hansen & Mikkel Flyverbom, *The Politics of Transparency and the Calibration of Knowledge in the Digital Age*, 22 ORG. 872, 889 (2015).

244. Gorwa & Ash, *supra* note 6, at 292–94.

245. *Id.* at 302.

content review team under Facebook's Community Standards.[246] Topics governed by the Community Standards include "adult nudity and sexual activity," "suicide and self-injury," "bullying and harassment," and "fake accounts" among others.[247] The most recent Community Standards Enforcement Report states that 98.8% of "suicide and self-injury" content they took action on (including but not limited to removal) was identified by Facebook (referred to as proactively detected) and the remainder was reported by users.[248] Given the nuanced and subjective judgments required to apply the suicide and self-injury policy that, for example, prohibits posting "content that focuses on depiction of ribs, collar bones, thigh gaps, hips, concave stomach, or protruding spine or scapula when shared together with terms associated with eating disorders" but allows those same images "in a recovery context," and similarly prohibits "content that depicts graphic self-injury imagery" but allows "older instances of self-harm such as healed cuts or other non-graphic self-injury imagery in a self-injury, suicide or recovery context,"[249] it is hard to know what this statistic means beyond the fact that Facebook believes its employees and algorithms are well-tuned to their policy.[250]

The reports, moreover, make it difficult to interpret differences in performance across content and over time. For example, evaluating within and across policy areas and over time is complicated by varying definitions of

246. Sarah Perez, *Facebook's New Transparency Report Now Includes Data on Takedowns of 'Bad' Content, Including Hate Speech*, TECHCRUNCH (May 15, 2018), https://techcrunch.com/2018/05/15/facebooks-new-transparency-report-now-includes-data-on-takedowns-of-bad-content-including-hate-speech/ (discussing a section of the report labeled the "Community Standards Enforcement Report").

247. *Community Standards Enforcement Report*, TRANSPARENCY CENTER, https://transparency.fb.com/data/community-standards-enforcement/ (Last Visited May 10, 2022)

248. *Community Standards Enforcement Report: Suicide and Self-Injury*, TRANSPARENCY CENTER, https://transparency.fb.com/data/community-standards-enforcement/suicide-and-self-injury/facebook/ (Last Visited May 10, 2022)

249. *Facebook Community Standards: Suicide and Self Injury Policy*, TRANSPARENCY CENTER, https://transparency.fb.com/policies/community-standards/suicide-self-injury/ (Last visited May 10, 2022).

250. For example, did Facebook's identification of material stop posting—meaning that users had no opportunity to catch it? Did such identification rely on automated tools? Facebook reports that in Q4 2021 only 200 items of content actioned under the suicide and self-injury policy were appealed resulting in 50 restorations. They further report that they independently restored 95,200 pieces of content of their own accord. This provides limited information on how the policy is applied and what content it effected. *Facebook Community Standards: Suicide and Self Injury Policy*, TRANSPARENCY CENTER,https://transparency.fb.com/policies/community-standards/suicide-self-injury/ (Last visited May 10, 2022).

"piece of content."[251] More information is necessary to understand the meaning of these numbers and what effect they might have on the public.

The voluntary nature and lack of standardization undermine the substantive impact and utility of these symbolic structures.[252] Transparency reports share some common features—they generally provide quantitative information on legal requests received on a country-by-country basis, sometimes broken down by issue, and they often provide information on the proportion of requests with which the firm complied. Yet despite efforts to promote standardization,[253] firms have not developed shared definitions or formats.[254] When information is missing, it can be unclear whether the company deems it insignificant, is legally constrained from revealing it,[255] or is intentionally withholding it.[256]

Transparency reports, moreover, are not merely partial; they are strategically so. Transparency reports "obfuscat[e] and redirect from more

251. *How Meta Improves: Content Actioned*, TRANSPARENCY CENTER, https://transparency.fb.com/policies/improving/content-actioned-metric/ (last visited May 10, 2022).

252. For an overview of some of the variations in definitions and reporting categories and styles, see generally LIZ WOOLERY, RYAN H. BUDISH, LEVIN BANKSTON, THE TRANSPARENCY REPORTING TOOLKIT: BEST PRACTICES FOR REPORTING ON US GOVERNMENT REQUESTS FOR USER INFORMATION (2016).

253. *Id.*; *Transparency Reporting Guidelines*, *Release*, GOV'T. OF CANADA (June 30, 2015), https://www.ic.gc.ca/eic/site/smt-gst.nsf/eng/sf11057.html.

254. Independent organizations have stepped in to press companies to address the limitations of their transparency reports. Ranking Digital Rights publishes a yearly ranking of select companies' policies relating to freedom of expression and user privacy. Its index provides an important benchmark supporting comparative analysis across companies and over time. The initial effort to provide transparency around DMCA complaints, chillingeffects.org, is now Lumen and collects and analyzes a broader range of content removal requests from around the globe providing a source of independent data on content removals and their impacts. LUMEN, *supra* note 69.

255. *See* Losey, *supra* note 63, at 3454–55 (describing various jurisdictions' limits on disclosing law enforcement requests).

256. *See, e.g.*, Rebecca Rose, *Google on Its Own Transparency Report: This is Not Good Enough*, THE ATLANTIC (Nov. 14, 2013), https://www.theatlantic.com/technology/archive/2013/11/google-on-its-own-transparency-report-this-is-not-good-enough/281473/ (discussing Google's 2013 report, in which it noted that it had received more requests from the government than ever before, but could not share data about all of the government requests due to the Department of Justice's contention that requests related to national security must remain private); Sam Shead, *TikTok Transparency Report shows it removed 49 million videos and had 500 government requests*, CNBC (July 9, 2020), https://www.cnbc.com/2020/07/09/tiktok-transparency-report.html (discussing TikTok's 2020 transparency report which had no data related to use of the app in China, where its parent company is based). The app goes by a different name in China, but it is unclear whether they intend to release information about that app in an additional report.

substantive and fundamental questions about the concentration of power, substantial policies and actions of technology behemoths."[257] The reports emphasize quantitative data, which is useful for some forms of analysis but lack the detail necessary to review and understand the legitimacy of legal demands and disclosures.[258]

Strategically partial reports limit more meaningful analysis. Statistics about the performance of machine learning tools, focused on accuracy and prediction, gloss over the human toll of content moderation failures. This emphasis on quantitative-transparency measures as the way to understand and evaluate content moderation skews public debates. All false labels are consequential but, in content moderation, some are far more so than others. For example, the recent failure of Facebook to remove the 3,000-member group calling themselves the Kenosha Guard that organized the armed response to the unrest in Kenosha, during which Kyle Rittenhouse shot three people, killing two, is a significant failure even if it was part of a quantitatively small number of pages that should have been removed.[259]

Thus, by producing statistics, rather than case studies or the more substantial records found in the Chilling Effects and Lumen database,[260] these

257. Monika Zalnieriute, *"Transparency Washing" in the Digital Age: A Corporate Agenda of Procedural Fetishism*, 8.1 CRITICAL ANALYSIS OF L. 139, 139–53 (2021).

258. *See, e.g.*, Losey, *supra* note 65, at 3454 (finding that Facebook, Google, LinkedIn, Microsoft, Tumblr, Twitter, Verizon, and Yahoo were not reporting the legal processes used by non-U.S. governments to access user data).

259. A Tuesday afternoon post read, "Any patriots willing to take up arms and defend our city tonight from the evil thugs?" Despite multiple reports under Facebook's terms of service by users, the account was left up until after the shooting. There are multiple provisions of Facebook's terms of service that could have led to the Kenosha Guard page being removed. Examples include provisions against Incitement to violence which bars "[s]tatements of intent or advocacy, calls to action, or aspirational or conditional statements to bring weapons to locations, including but not limited to places of worship, educational facilities or polling places (or encouraging others to do the same)." *See Violence and Incitement*, META https://www.facebook.com/communitystandards/credible_violence (last visited Mar. 24, 2022); *see also Dangerous Individuals and Organizations*, META, https://www.facebook.com/communitystandards/dangerous_individuals_organizations (last visited Apr. 20, 2022) (provisions limiting posts by "Dangerous Individuals and Organizations"); These provisions were updated on August 19, 2020 to more explicitly remove accounts hosting discussions of potential violence adopted to address militia organizations and others, *An Update to How We Address Movements and Organizations Tied to Violence*, META (Aug. 19, 2020), https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence/, and eventually served as the basis for the group's removal. *See* Russell Brandom, *Facebook takes down 'call to arms' event after two shot dead in Kenosha*, THE VERGE (Aug. 26, 2020), https://www.theverge.com/2020/8/26/21402571/kenosha-guard-shooting-facebook-deplatforming-militia-violence.

260. *See supra* note 71.

reports shape understandings of transparency and the way "success" and "failure" are understood and measured in content moderation. It suggests we take a thousand-foot view—a bean counter's view—asking "how much," while directing attention away from qualitative measurements centered on the substantive impact of content moderation practices. The statistics reported provide no insight into concerns about inconsistent or biased applications or outcomes of rules. In this way, transparency reports perform a sort of "transparency washing,"[261] directing analysis in a certain way to answer certain questions, while at the same time obscuring data necessary to answer the substantive normative questions that undermine a platforms' ability to legitimately perform the application subfunction.

By design, transparency reports provide a limited and skewed glimpse into corporate practice. The reports provide a very partial picture of content removals and strategically divert attention away from the substance of corporate content moderation policies, the procedures for removals pursuant to them, and the substantive impact of those removals on speakers and listeners. This redirection of attention is consistent with platforms' overall interest in limiting the extent to which the public and policymakers were aware of their vast and largely unchecked moderation activities, particularly those involving greater human review.[262]

Through the lens of the resolution or application subfunctions, then, transparency reports do not fill the gaps that undermine platforms' ability to perform it legitimately. The statistics do not provide substantive information about the who and why of content moderation, nor do they help answer important questions about procedural regularity and consistent, impartial application of rules. Even with respect to the narrow substantive task of holding governments to account, researchers believe they are ineffective.[263] Industry insiders have mixed views on their substantive impact. Some find that some governments were emboldened by data about the number and success of other governments' requests for data and content removal. Others reported that that transparency about government requests improved government

261. Zalnieriute, *supra* note 257, at 139.

262. *See* TARLETON GILLESPIE, CUSTODIANS OF THE INTERNET: PLATFORMS, CONTENT MODERATION, AND THE HIDDEN DECISIONS THAT SHAPE SOCIAL MEDIA *passim* (2018) (describing social media platforms efforts to limit public awareness of content moderation through strategies including emphasizing the role of algorithms in content moderation, and keeping human workers in the shadows).

263. Losey, *supra* note 65, at 3456 (concluding data in transparency teports are insufficient for oversight or debate and urging companies to increase the granularity of reported data and provide greater specificity about legal processes governments use).

adherence to rule of law, evident in narrower requests with clear legal basis and greater procedural regularity.

Transparency reports could, if designed differently, address a systemic failing in content moderation systems and practices: the lack of visibility into what is removed, on what basis, and at whose request. In doing so, they could provide a means of holding companies accountable for maintaining standards of procedural regularity, consistent application of rules, and fair and equal treatment—thereby helping the public hold governments to account for speech removal practices. In their current design, however, they operate more as a symbol than substantive protection.

The strategic role that transparency reports play underscore their largely, though not purely, symbolic nature. At their worst, they do not establish the trustworthiness or legitimacy of firms to perform the moderating function[264] but rather operate as transparency theatre, diverting attention to other actors and away from platforms. Yet, they are a symbolic legal structure with promise. Their evolution from information about government and third-party removals to information about removals under corporate rules is promising. However, moving from performative transparency to symbolic legal structures that align content moderation practices with democratic norms requires access to qualitative data and more detailed statistical data.

While each of these three structures—transparency reports, Google's Advisory Council, and Facebook's Oversight Board—respond to some extent to the values at stake in the relevant subfunction at issue, none robustly integrates the competencies necessary to render the exercise of platform power substantively and procedurally legitimate. Analyzing these structures in relation to content moderation subfunctions exposes their limitations and, as we develop in the next Section, offer a tool for reconfiguring content moderation to yield more legitimate distributed content governance frameworks.

B. THE CONSTRUCTIVE TURN: USING A FUNCTIONAL FRAMEWORK TO CONFIGURE CONTENT MODERATION

As its application to these examples demonstrates, a focus on subfunctions and constraints helps surface the ways that content moderation has been (and can be) organized and identifies the potential social and political implications of those design choices. In doing so, it clarifies why particular content moderation regimes draw particular, and particularly vociferous, objections. More importantly, consistent with the aims of the New Governance research

264. *Id.*

described above,[265] the functional typology can assist regulators and other stakeholders in constructing content moderation regimes to align competencies with subfunctions and to couple them with constraints that further align delegations with public values. It provides a playbook, or a set of design patterns,[266] to assist us in coupling allocations with constraints to build content moderation systems that adhere to public values, despite being largely composed of private entities. Below we walk through an example that illustrates the benefits of the framework in making this constructive turn.

1. *Leveraging Competencies and Addressing Democratic Deficits: Reimagining the Global Internet Forum to Counter Terrorism*

To meet their commitment under the EU Code of Conduct on Countering Illegal Hate Speech Online, Facebook, Google, Microsoft, and Twitter committed to, among other things, create the Global Internet Forum to Counter Terrorism (GIFCT). A key activity of the GIFCT is maintaining a shared hash database of terrorist content.[267] While initially developed by the four companies, it is now used by thirteen companies.

On the surface, this shared hash database might appear to be a reasonable extension of the approach pioneered with the NCMEC CSAM hash database described above. It too provides a common resource to support the *locating* subfunction and by doing so, reduce the cost—financial and human—of multiple companies, and individuals within them, having to review and *identify* some of the most gruesome and disturbing content on the web. Further, by pooling the knowledge of identified terrorist content and automating the locating function across different platforms, the GIFCT database can speed up the removal of content valorizing violence. Coordinated multi-platform action is considered exceedingly important to stop the spread of violent content that can spread virally, build visibility and support for terrorist ideologies, from white supremacism to religious extremism, and fuel copycat actions.

---

265. *See supra* text accompanying notes 84–89.

266. For an alternative way to think about the construction of a subset of content moderation policies, see generally DAPHNE KELLER, BUILD YOUR OWN INTERMEDIARY LIABILITY LAW: A KIT FOR POLICY WONKS OF ALL AGES (2019).

267. In 2016, Facebook, Microsoft, Twitter, and YouTube announced they would work together to create a shared industry database of online terrorist content. *Partnering to Help Curb Spread of Online Terrorist Content*, FACEBOOK (Dec. 5, 2016), https://about.fb.com/news/2016/12/partnering-to-help-curb-spread-of-online-terrorist-content. This database later became part of the Global Internet Forum to Counter Terrorism. *About*, GLOB. INTERNET F. TO COUNTER TERRORISM, https://gifct.org/about (last visited Mar. 24, 2022).

Despite these similarities to the NCMEC hash database, the GIFTC database has been roundly criticized. The variety of concerns and objections are captured in a letter sent to the EU objecting to the then-draft EU proposal for regulating the dissemination of terrorist content online by an impressive and large group of the human rights organizations, journalists associations, and researchers.[268] They noted that the definition of terrorist content is unstable, unshared, and subject to exceptions that protect important public accountability functions including news reporting and human rights documentation.[269] Absent a shared and stable definition, using a set of distributive identified materials to drive removals across platforms and jurisdictions raised substantial concerns. In addition, given the importance of contextual evaluation, they raised concerns about the use of automated content moderation tools, in particular upload filters.[270] Given the substantial risk of over removal, they objected to the profound lack of transparency and accuracy that generally attends automated decision-making.[271] They argued that "because it is impossible for automated tools to consistently differentiate activism, counter-speech, and satire about terrorism from content considered terrorism itself, increased automation will ultimately result in the removal of

268. Access Now et al., *Joint Letter to EU Parliament: Vote Against Proposed Terrorist Content Online Regulation* (Mar. 25, 2021), https://www.hrw.org/news/2021/03/25/joint-letter-eu-parliament-vote-against-proposed-terrorist-content-online#. The regulation, adopted in June 2021, underwent substantial amendments, several of which addressed concerns raised by the group. Regulation (EU) 2021/784 of the European Parliament and of the Council of 29 April 2021 on addressing the dissemination of terrorist content online, 2021 O.J. (L 172) 79. While it maintains the one-hour time period for hosting service providers to remove terrorist content upon notice by a competent authority, it allows for delay for "objectively justifiable technical or operational reasons," and it requires annual reports by governments and platforms about removals, user notification of content removal determinations, and appeals processes, uses the definition of terrorist content from the existing Directive (EU) 2017/541 of the European Parliament and of the Council of 15 March 2017 on combating terrorism, 2017 O.J. (L 88) 6, and includes exceptions for content distributed for journalistic, research, or artistic purposes. *See* Katrien Luyten, *Addressing the dissemination of terrorist content online*, EUR. PARLIAMENTARY RSCH. SERV. (July 15, 2021), https://www.europarl.europa.eu/thinktank/en/document/EPRS_BRI(2020)649326 (describing legislative process, including stakeholder engagements, and final proposal); *Terrorist Content Online*, (Apr. 2021), https://ec.europa.eu/home-affairs/document/download/506de61d-c53f-489f-a9e3-e16e78793e81_en. Relatedly, the French Constitutional Council, which reviews the constitutionality of legislation, struck down key provisions of the French Law on Countering Online Hatred, because they required platforms to make decisions about the legality of content without judicial involvement. Aurelien Breeden, *French Court Strikes Down Most of Online Hate Speech Law*, N.Y. TIMES (June 18, 2020), https://www.nytimes.com/2020/06/18/world/europe/france-internet-hate-speech-regulation.html.

269. Access Now et al., *supra* note 267.

270. *Id.*

271. *Id.*

legal content like news content and content about discriminatory treatment of minorities and underrepresented groups."[272] They objected to the lack of procedural constraints, writing that "[t]he lack of judicial oversight is a severe risk to freedom of expression, assembly, association, religion and access to information."[273] Finally, given the lack of agreement on definitions, they raised concerns about a system that would paper over conflicting local laws and norms and allow "one Member State [to] extend its enforcement jurisdiction beyond its territory without prior judicial review and consideration for the rights of individuals in the affected jurisdictions."[274]

These critiques elucidate how subject-matter-specific attributes of policy definition and application combine with the specific allocations of subfunctions in the GIFCT to undermine its substantive and procedural legitimacy.

The CSAM database is run and populated by NCMEC, a nonprofit entity that holds expertise in child sexual exploitation and is independent from the commercial platforms.[275] NCMEC has particular operational capacity—it is funded and legally authorized to do things other entities cannot—and reflects the interests of victims and the common interest in eradicating both CSAM and the underlying activity it captures.[276] NCMEC is tasked with applying the definition to content it receives to determine whether it should be added to the CSAM database. The hashes in that database, already determined to meet the definition of child sexual abuse material, are then used to identify and locate images that match (or nearly match) them on participating services. The definition of CSAM is established in public law in the United States, a definition that is consistent and at times identical to those in other countries. However, as described above, to avoid edge cases, NCMEC has chosen to limit the CSAM in the hash database to images considered the worst of the worst. CSAM material is illegal without exception. There is no journalistic, research, or other exceptions that make its storage, viewing, or distribution legal. For these reasons cultural and contextual cues are largely irrelevant to the application of the rule, limiting the potential gap between identification and application. To the extent that content can be classified on its four corners, it is deemed to meet the definition.

For these reasons, identification of this more limited set of CSAM should be both relatively straightforward and relatively consistent across platforms—

272. *Id.*
273. *Id.*
274. *Id.*
275. *See supra* text accompanying notes 134–139.
276. *Id.*

informed by shared law rather than by variable corporate policy. In this context, relying on a shared set of identified materials should pose little additional risk of over removal because identified content is likely to be subject to moderation under the rules. Finally, technology is used only to locate content that matches human-labeled content, not to independently identify new CSAM to which the definition has not yet been applied. As a system, this functional arrangement aligns subfunctions with competencies and builds in constraints (the narrower set of images included) to address potential risks and leverages technical actors (automation) in a discrete manner.

The configuration of the GIFCT Shared Industry Hash Database differs from the CSAM structure in fundamental ways. First, there is no shared public definition to guide the GIFCT's actions. There is no globally accepted agreement on the definition of terrorist content, and company standards vary. The lack of a shared and stable definition raises concerns about the appropriateness of pooling content identified across platforms and, importantly, across borders and targeting it for removal. In addition, unlike CSAM, there is a recognized need for exceptions. For example, terrorist content may be posted, shared, and stored to document human rights violations or to report on them. Applying policies to discrete pieces of content that may be embedded in larger works—archives of abuse or news stories, for example—makes judgments outside the context important to interpret the meaning of content in this area. Thus, with respect to terrorist content there is likely to be a substantial gap between content identified as potentially actionable and content found to be so after the rules are applied.

Second, the GIFCT is an industry consortium, not a non-profit with specific expertise, raising concerns about its independence, representativeness, and expertise. Secrecy around the structure of the work, including who identifies content as appropriate for inclusion in the database, and how GIFCT deals with the distinct policies of different companies, have furthered concerns about independence and expertise.

Finally, while technical actors are used to locate content that matches material previously identified for removal in ways similar to the CSAM database, the lack of clarity as to who determines what is in the database and what criteria are used to determine inclusion make this matching activity far more problematic. The technical means are the same, but the impact is decidedly different. In a partial effort to address this, initial corporate statements emphasized that matches against the database would not result in automatic removal; however, recent statements and reports indicate this is no

longer true: companies have backed away from their initial constraint of platform specific human review before removal.[277]

A functional approach suggests several ways subfunctions could be reconfigured to bring greater legitimacy into the content moderation regime for terrorist content. First, with regards to the task of defining, governments could establish a stable public agreement on the definition of terrorism content.[278] The GIFCT, like NCMEC, moreover, could publicly agree to use the shared identifying and locating resource of the database for a narrower subset of terrorist content. Second, the GIFCT could create a representative body of independent experts to develop implementation guidelines for this narrower set of content. Third, the GIFCT could create an independent body to make determinations under this definition, limiting the ability of any platform to poison the system with poor evaluations and ensuring both expertise and independence in these decisions. This is of particular importance where liability regimes create risks to platforms—legal or other—that may incentivize over removal.

Finally, given the importance of context to the proper identification of terrorist content, the system should build in additional constraints, both technical and human. On the technical side, locating content could be constrained to ensure that material is not removed from known news sites or human rights archives, using things such as domain-level blocking. Unlike in the CSAM arena, where context, including the site on which content is located, is unimportant to whether it can be regulated, when dealing with terrorist content, context informs the legitimate application of the definition. The exception for content distributed for journalistic, research, or artistic purposes included in the EU regulation on the dissemination of terrorist content online underscores the potential importance of location and other factors that contribute to an understanding of purpose.[279] In addition, the content of the database could be subject to review and audit by independent experts and coupled with broader transparency reports, like those envisioned by the Santa Clara Principles,[280] to ensure automation does not come at the cost of speech

277. *See* BSR, 2021. "Human Rights Assessment: Global Internet Forum to Counter Terrorism" at 41.

278. *See id.* at 33 (concluding, after an assessment of GIFCT and consultation with stakeholders, that the task of creating a shared definition of terrorist and violent extremist content "properly resides with governments").

279. Regulation (EU) 2021/784 art. 1, ¶ 3, 2021 O.J. (L 172) 79, 89.

280. *Santa Clara Principles 1.0*, THE SANTA CLARA PRINCIPLES ON TRANSPARENCY & ACCOUNTABILITY IN CONTENT MODERATION, https://santaclaraprinciples.org/scp1/ (last visited Mar. 24, 2022). Business for Social Responsibility recommends that GIFCT figures out

that is particularly valuable, such as documenting human rights abuses and journalistic coverage of terrorist events.

Through reallocating and constraining subfunctions, the benefits of this shared identification and location resource could be realized in a manner that is substantively and procedurally legitimate.

V. CONCLUSION

In his forward-looking 1998 Article, *Lex Informatica: The Formulation of Information Policy Rules Through Technology*, to which this Issue is dedicated, Joel Reidenberg foresaw many of the elements that would come to characterize the regulatory environment of information infrastructures. The trans-jurisdictional scope of networks, he wrote, would decentralize the role of traditional government regulation in content governance, leading to fragmentation and confusion in rules about information flows. Sovereign laws could still wield some influence, but other forms of rulemaking and enforcement, including technical solutions and system design choices framed by private actors and those who designed technology for them, would shape policy decisions.

Our development of a functional framework for understanding, assessing, and constructing content moderation reflects Reidenberg's challenge to policymakers to "understand, consciously recognize, and encourage" the actual workings of these distributed forms of governance. Its identification of subfunctions supports a rigorous analysis of the way that content moderation actually functions on the ground, the concrete choices available regarding allocations of discrete subfunctions to different public or private actors, and how to leverage different actors' capacities and competencies. It surfaces the normative implications of different content moderation configurations and thereby facilitates an assessment of how such allocations can be constrained—either through processes or through limits on the use of automation—in ways that address those normative concerns.

Through this lens, a functional framework both offers a means for critically assessing the way various content moderation regimes allocate and constrain various subfunctions and generates constructive insights regarding the structuring of the content moderation function in new contexts. Its focus on the relevant content moderation subfunction involved in three structures that private platforms have adopted in response to legitimacy critiques—Google's Advisory Council on the Right to be Forgotten, Facebook's Oversight Board, and the use of transparency reports—surfaces the particular normative

how to enable annual, publicly reported, third-party reviews of the hash-sharing database. BSR, 2021. "Human Rights Assessment: Global Internet Forum to Counter Terrorism" at 44.

concerns implicated by the particular task and the governance competencies needed to address them. It thus enables a granular analysis of ways that the task is structured—to whom the subfunction is allocated, and how its exercise is constrained—and points to the ways that each fall short.

Looking forward, the functional framework permits a proactive analysis into how subfunctions in content moderation regimes might appropriately be structured, through law and private action, to promote legitimate governance systems in the future. It thus responds to Reidenberg's charge, providing a framework that regulators and others can use to construct distributed content governance regimes that restrain the power of platforms to pursue narrow self-interests while leveraging their capacity and expertise, along with that of other stakeholders, to advance the public interest.